THE UNIVERSITY *of* EDINBURGH

Joint Modelling of Longitudinal and Survival Data for Dynamic Prediction in Credit-Related Applications

Víctor H. Medina-Olivares

Doctor of Philosophy

The University of Edinburgh

2022

Lay Summary

The money lending process considers several assessments that support the lender's decision-making throughout the credit lifecycle. In the beginning, for example, lenders evaluate whether an applicant will be able to repay the loan in the future or, equivalently, how prone they are to the event of default. Once the credit is granted, lenders, in light of new data gathered about the performance of their borrowers, focus on updating these assessments or determining the likelihood of other events that affect the expected return of the loans (e.g., prepayment of the total amount owed). This monitoring is supported by mathematical models that estimate how probable the event of interest will occur in a future period. Therefore, more reliable models sustain competitiveness and good management of the associated risks. In this thesis, we are interested in dynamically predicting when credit-related events occur. Dynamic, in this context, refers to a setting where we can update predictions given the newly collected data.

Survival models are widely used in the credit risk literature to make dynamic predictions. These models answer, for example, how likely a particular borrower will default for different time horizons, given the data available. These data include variables such as loan repayment history, the use of other credit products and even related to economic conditions. However, the sound way to build these models must question whether these variables would eventually be influenced by the occurrence of the event or not. For example, we would not expect a borrower default event to influence the future path of the unemployment rate. However, a borrower's loan repayment path would be strictly related to its default. These latter sorts of variables are called endogenous variables. We must address this mutual evolution when incorporating these variables in a survival analysis context. The commonly used techniques do not handle this joint evolution and could harm the quality of the prediction.

The thesis studies relatively new models in the credit risk literature called joint models of longitudinal and survival data. In their standard version, these models are formed of two sub-models, one survival and the other related to the trajectory of the endogenous variable. Furthermore, the approach postulates that these two sub-models are connected. In this way, we can address the mutual evolution, employ what the model has estimated with the data collected up to a given point, and project this evolution for predictive purposes.

These models, like survival models, have their origins in medical research. In medical applications, the interest can be, for example, the time to the appearance of a disease and how it relates to the evolution of biomarkers. Although the analogy to credit-related applications is straightforward, there are essential distinctions. This thesis introduces several innovations for a more appropriate joint model approach.

In particular, we modify the standard formulation so that predictions and evaluation metrics are consistent with the monthly frequency of loan reporting. Moreover, we present a methodology that handles variables whose values at a given moment can be partly described by their values observed in the past. In addition, we reformulate the joint models to estimate them more quickly than commonly used techniques. That facilitates exploring appealing designs in line with the credit context, such as leveraging large dataset environments, including more than one endogenous variable or considering the borrower neighbourhood effect in the model. Finally, we propose new methodologies to estimate dynamic predictions accurately and efficient ways to compare different models.

We apply joint modelling approaches to default and prepayment events in mortgage and consumer loan portfolios. Empirical results reveal that these models are viable in credit-related applications and can achieve better predictions than standard survival models.

Acknowledgements

I am most grateful to my supervisors, Raffaella, Finn and Jonathan, for openly sharing their experience and knowledge and being engaged throughout this journey. Without their feedback, this could not have been possible. I also value the anonymous reviewers for their helpful comments in the early versions of the papers that form part of this thesis. Finally, I thank the University of Edinburgh for recognising my research and supporting it.

On a more personal note, reaching this step implies plenty of work and, above all, a lot of luck in having the backing of wonderful people. I am fortunate to have grown up in a beautiful family environment, the fruit of generations that learned how to overcome their struggles and shortcomings in a country where opportunities do not abound. Grandpa, I remember you more than ever, conveying your hopeful thoughts. Thank you for making me a better version of myself.

I want to thank my parents for their immense love and support throughout my years. Their enduring message, "the greatest inheritance we can leave you is your education", sounds louder than ever. Thanks to my two sisters for caring and being an active part of my life. Likewise, thanks to my uncles, Willy and Viviana, for supporting me at the beginning of my formative stage.

Last but not least, I want to acknowledge Martha, the love of my life, for her ongoing encouragement and generous love and for tolerating my absence during this period.

To Lili and Martha.

Abstract

Lenders monitor their borrowers over time, allowing them to dynamically predict the probability of an event of interest, such as default. The widely used survival models focus on when the event happens and can handle time-varying covariates (TVCs) and censored observations. However, an issue little addressed in the literature is that the model specification and the predictive framework depend on the type of TVC included. TVCs can be either exogenous or endogenous to the survival time. Exogenous are those whose future paths are not affected by the event's occurrence, such as macroeconomic variables. Endogenous, on the contrary, are those whose paths are influenced by the survival status. An example of the latter would be the unpaid principal balance when the event is the default.

This thesis explores new mathematical models in credit-related applications, known as joint models of longitudinal and survival data. Initially developed in medical research, these models, in their standard version, are formed by two sub-models, one for the survival process and the other for the endogenous TVC (also named longitudinal outcome in this context). A latent structure links the sub-models, commonly in the form of random effects. Joint models have two advantages compared to survival models. First, they allow us to handle possible endogeneities in the TVCs. Second, by jointly modelling both processes, they offer us a dynamic prediction framework that incorporates their mutual evolution.

We propose a series of innovations to make the approach appropriate to credit-related applications. These innovations relate to the nature of survival time, the specific evolution of the TVCs, ways to scale the technique to large datasets and how to leverage the available data in the modelling framework.

In concrete, we adapt the formulation of the joint models and their performance metrics to the discrete nature of the loan data. In addition, we include autoregressive terms in the TVC specification to address observed serial correlation and enhance predictive capability. Moreover, we can study more complex specifications with larger datasets by reformulating the approach within the INLA framework, a fast and accurate algorithm for Bayesian inference. Among these specifications are the joint models with more than one TVC and the joint model that leverages geographical information to include spatial and spatio-temporal effects in the hazard function. We also introduce a more accurate way to estimate

individual survival predictions using the Laplace method. Finally, to compare different models, we propose a computationally efficient implementation of the cross-entropy estimate of the posterior predictive conditional density that uses the estimates obtained in the inference step.

We apply joint models to predict the time to credit events in the following three settings: default in US mortgages, full prepayment in a German consumer loan portfolio, and full prepayment in US mortgages. The main empirical results show that the autoregressive terms in the joint model let us achieve better discrimination performance, the predictive ability is significantly enhanced compared to survival models when more TVCs are considered, and the inclusion of spatial effects consistently leads to better data representation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

1.1.1 Credit context

Credit is the ability to receive money with the understanding that it will be paid later. It is a vital source of liquidity for financial transactions in the public and private sectors and an essential means for economic growth (Beck et al., 2000). This way of funding permits, among other things, companies' daily operations, long-term investments, owning a house or even starting a business. Conceptually, credit has been present in our lives since the beginning of civilisation (Doumpos et al., 2019, Ch.1). However, in the last few decades, we have witnessed enormous changes in the credit market concerning its volume, type of credits, channels to provide it, and how it is regulated.

Table 1.1 shows data from the Bank of International Settlements (BIS)[1] reflecting credit expansion in the private non-financial sector (non-financial corporations and households) over the last twenty years for China, the Euro area and the United States (figures in domestic currency billions). For example, in 2000, China's total credit amounted to 10,957 billion Yuan, corresponding to 109% of its gross domestic product (GDP). In 2020, this amount increased to 224,901 billion Yuan, equivalent to 223% of GDP. Without such a spectacular but undoubtedly significant expansion, the Euro area and the United States, over the same period, have increased credit to non-financial corporations and households

[1] `https://stats.bis.org/`

by 49 and 30% relative to GDP, respectively.

	2000		2020	
	Total credit	% of GDP	Total credit	% of GDP
China	10,957	109	224,901	223
Euro area	8,993	126	19,553	175
United States	13,885	135	34,429	165

Table 1.1: Total credit to the private non-financial sector in domestic currency billions and as a percentage of GDP.

Moreover, like many other markets, financial services have been stimulated by technology and innovation. Financial Technology, known as FinTech, is an expanding industry that has challenged incumbent financial institutions through the use of technology. This industry has gained prominence in the market by allowing broader and more efficient access to the end customer in various lines of business such as wealth management, investments, trading, payment methods and lending. For example, considering the unsecured personal loan market, FinTech companies in the United States went from a market share of 22% in 2015 to 49% in 2019[2]. In response, traditional lenders have had to adopt similar technology, further boosting competitiveness, particularly in the lending business. All of this has influenced the extension of the service offering and the reach to new customer segments.

Credit expansion, technological advances, and an increasingly interconnected financial world have posited challenges in managing and supervising financial risks. Recognising this, the main standard-setter for the prudential regulation of banks, the Basel Committee on Banking Supervision (BCBS), has introduced a series of supervision recommendations known as the Basel Accords[3]. These Accords have evolved to adjust to the growing sophistication of the banking industry and enhance financial stability. The first Basel Accord, Basel I, was presented in 1988 and set guidelines for capital requirements focused mainly on credit risk. However, these guidelines were oversimplified and consisted of minimum capital requirements for banks based on a weighting scheme of their assets. In 2004, acknowledging the limitations of Basel I, the Committee presented Basel II. This

[2] https://www.experian.com/blogs/insights/2019/09/fintech-vs-traditional-fis-latest-trends-personal-loans/

[3] https://www.bis.org/basel_framework/

more refined Accord included capital requirements to address not only credit risk but also market and operational risks. Yet, these new rules were inadequate to prevent the subprime financial crisis that started in 2007, and a new Accord was developed. Basel III, expected to be in place on the 1st of January 2023[4], aims to strengthen the Basel II capital requirements and introduce new requirements on assets liquidity and funding stability.

Credit risk management is a complex and constantly improving process with regulatory, technical and methodological challenges. Furthermore, its scope is not only relevant for the particular lender but also for the proper functioning of the financial system. For a management process to work correctly, the timely identification, measurement, monitoring, reporting, control and mitigation of the risks associated with the entire credit life cycle is essential. To this extent, the risks can be originated from any potential that the borrower will fail to meet future obligations by the terms agreed in the contract. Yet, the crucial point is that the lender is ultimately the one who accepts who is subject to credit and under what contractual terms. Therefore, how much risk is acceptable to the company strictly depends on the risk-reward ratio they can assess.

A clear example of the risks associated with credit is when the borrower, for whatever reason, does not repay the loan (BCBS, 2000). Under these circumstances, the consequences for the company's profits are clear. Thus, before loans are granted, potential borrowers are evaluated based on their ability to repay in the future. Indeed, much of the effort in credit risk management goes into creating these evaluation protocols, commonly supported by statistical models. However, the intrinsic nature of reckoning an event that has not yet occurred means that even the best protocol will eventually fail. Knowing this, lenders set loan provisions for potential losses, and the size of these provisions is in line with what they measure as expected losses over a fixed time horizon (e.g., 12 months). Getting the right amount of provision is essential. If the evaluation is too conservative, meaning loan loss provisions above the actual losses, rule out the option of profiting from the overestimated amount (e.g., granting more credit). On the contrary, if the amount is underestimated, the losses are balanced with the bank's capital. If the latter is insufficient, it can directly affect banks' intermediary role

[4]Initially, it was scheduled to be gradually implemented between 2013 and 2015. Then, it has been repeatedly rescheduled. The last postponement from January 1, 2022, to the following year is due to the pandemic.

in linking depositors and borrowers.

While credit default is given the importance it deserves, lenders also recognise that it is not the only risk associated with the credit life cycle. Prepayment risk, for example, is the risk that the borrower repays the loan before the term stipulated in the contract. The reasons for prepayment can be driven by better market conditions to refinance the loan, changes in the collateral value, and renegotiation of new loans with the same lender, among others (Consalvi and Scotto di Freca, 2010). When prepayments are made, the lender stops receiving scheduled cash inflows, directly affecting earnings by reducing future interest flows. However, prepayments can trigger other effects, such as maturity mismatch. Loan portfolios are financed through bank liabilities such as deposits and bonds. Both assets and liabilities have their respective maturities, and the lender, through different mechanisms, tries to match future inflows and outflows. As a result, prepayments produce a mismatch in maturities, and if they are not anticipated on time, they can induce over-or under-financing.

We mentioned the risk of default and prepayment. However, despite being fundamental when defining optimal credit decision policies (Ma et al., 2010), these risks are not the only ones for which lenders must protect themselves. Rather, we describe them because, as we will see below, this thesis explores new approaches for credit-related applications, precisely motivated by the prediction of default and prepayment events. In particular, we build mathematical models, known as a credit scoring system, to serve as a technical tool to support decision-making.

1.1.2 Credit scoring systems

Analysing the risks in the credit industry comprises qualitative and quantitative extents. Quantitative analysis plays a significant role, in particular, in retail and small business lending, where decisions need to be made over many clients. To help decision-making, lenders build mathematical models to estimate the probability of the occurrence of the event of interest for each client. These models, known as credit scoring systems, have traditionally been used to predict customer default on specific products. However, today the use and purpose of these models are much broader and in line with the profit maximisation objectives of the lender. Some applications where these models are used apart from the prediction of default are in predicting early repayment, fraud, usage, retention, etc. (Thomas

et al., 2017). Depending on whether these models are aimed at new customers or existing ones, they are commonly referred to as an application or behavioural models, respectively (see Thomas, 2000, for a detailed discussion between these models).

These systems, whether application or behavioural, are built under different techniques, all with the common goal to describe and predict the credit phenomenon in question in the best possible way. Classification analysis, the most used technique in the industry, uses cross-sectional data to estimate the probability that the event of interest will or will not occur within a predefined time horizon. This fixed planning horizon is usually 12 months, as suggested by the BCBS (2004). The setting is then a binary classification problem where historical data are statistically analysed to extract the borrower covariates most beneficial to the prediction.

Several classification algorithms have been explored in this regard, with greater emphasis on predicting the default event. To mention a few, neural networks (Baesens et al., 2003; Khashman, 2010), classification trees (Arminger et al., 1997), support vector machines (Wang et al., 2005; Bellotti and Crook, 2009b), smoothing nonparametric methods (Liu et al., 2009), and more computationally expensive ones like deep neural networks, heterogeneous ensembles, among others (Lessmann et al., 2015; Dastile et al., 2020). Historically, the most used ones are discriminant analysis and logistic regression (Hand et al., 2001).

However, classification analysis approaches have some limitations. The most evident is the need to predefine a fixed planning period, restricting the analysis to other time horizons. Other problems can arise, for example, when the outcome period does not include a portion of the credit cycle where a significant part of the events occurs in the portfolio. In this case, we shall be incurring an underestimation of the risk in that portfolio. That happened, for instance, in the subprime crisis of 2007, where some credit scoring systems were built using the first year of the mortgages' performance. Yet, most of the defaults occurred between the second and third years when the interest rates increased compared to those at origination (Thomas et al., 2017).

In addition, the performance period in building these models is often too short to observe substantial variability of covariates over time, making it challenging to

include time-varying effects in the modelling framework, which has proven useful when forecasting (Bellotti and Crook, 2009a; Stepanova and Thomas, 2002; Dirick et al., 2019). Furthermore, the credit scoring modeller needs to decide how to treat limiting cases, such as where the borrower's records are not complete within this period due, for instance, to an early closure of the account; or when the event has not occurred by the end of the period, but the borrower has given clear signals that it will (e.g., when the credit is in arrears and the event is the default). Usually, these cases are removed from the training sample, discarding valuable information (Thomas, 2000).

Another widely used technique for building credit scoring systems is survival analysis (Thomas et al., 2017, Ch.5). Rather than predicting whether the event of interest will occur or not within a fixed planning horizon, this approach estimates when the event is likely to occur. That allows analysing the event under different time horizons and overcoming the mentioned issues. That is to say, survival methods permit the inclusion of any outcome period in the study, time-varying covariates in the predictor and incomplete/censored performance records without the need to discard valuable data. This thesis is in the context of survival analysis with the inclusion of time-varying covariates.

1.2 Motivation

Lenders periodically collect data, such as variables that describe the economic conditions, account-level data about the performance of their portfolios or updated characteristics of the borrowers. Which and how many of these time-varying covariates (TVCs) are tracked depends on the lender's capacity to pull and join different sources of information. However, some common records include macroeconomic variables (GDP, interest rate, unemployment rate, etc.), unpaid principal balance, scores from external companies such as Fair Isaac Corporation (FICO), number of credit products with the lender, arrears in instalments, income, and collateral evaluations, among others (Bellotti and Crook, 2014; Djeundje and Crook, 2018; Calabrese and Crook, 2020). Furthermore, when building a model for predictive purposes, one wants to use the most updated information. The attention is then on exploring ways to include TVCs in the modelling framework.

Survival analysis provides a framework that facilitates the inclusion of TVCs. The literature on survival credit analysis has shown that TVCs can improve the model's predictive performance (see Section 2.1.4). Nonetheless, the type of TVC included in the model dictates the estimation and predictive procedures, and only a few papers address these topics. TVCs can be linked to a specific account, such as unpaid principal balance, or not, such as macroeconomic variables. Those that are account-specific can be further categorised as exogenous or endogenous to the survival status. Exogenous TVCs are whose future paths are not affected by the event's occurrence but affect the outcome. On the contrary, endogenous are influenced by the survival status and thus carry direct information on the event timing (see Section 2.1.5 for further discussion between exogenous and endogenous TVCs). To this extent, all the TVCs that are not specific to the account, such as macroeconomic variables, are exogenous, and not all account-specific TVCs are endogenous to the survival time. An example of the latter could be the borrower's income when predicting the time to credit default. One can think that income, which is borrower-specific, can directly affect the time to default. However, the default will probably not influence the borrower's future income.

We can find survival models in credit risk literature that include, in addition to time-fixed covariates, only TVCs that are not specific to the account (Bellotti and Crook, 2009a; Dirick et al., 2019), only account-specific TVCs (Stepanova and Thomas, 2001; Crook and Bellotti, 2010) or both (Djeundje and Crook, 2019a; Calabrese and Crook, 2020). However, of those that include account-specific TVCs, the distinction between their endogenous nature is commonly overlooked, meaning there is no control for this matter. The statistical consequence of not controlling by endogeneity is incorporating estimation bias (Wulfsohn and Tsiatis, 1997; Tsiatis et al., 1995). Since we commonly do not know the true data generation process, one might think that bias measurement is impractical. Also, if models are built for predictive purposes, why should we bother measuring bias in the first place? Regardless, when we use account-specific TVCs in a survival model, standard techniques to make predictions are to carry forward the last available observations of the TVCs or lag their values to relate past information to future survival status. Both techniques have limitations, as we will see further in Section 2.1.6. One is that we are not relating synchronised observations between TVCs and survival paths, which could indeed affect predictive perform-

ance. Therefore, even if we focus on prediction and overlook potential endogeneity issues, the relationship commonly assumed between endogenous TVCs and the time to event appears suboptimal.

The primary motivation of this thesis comes from exploring a relatively new approach to the literature on credit analysis, namely the joint modelling of longitudinal and survival data (Elashoff et al., 2016; Fitzmaurice et al., 2008; Rizopoulos, 2012). This modelling approach allows us to control for potential endogeneities of the TVCs and offers a dynamic predictive framework that does not rely on lagging the TVCs or extrapolating the last observations. Instead, the approach takes advantage of the estimated association between the evolution of the endogenous TVCs and the survival path and casts this mutual evolution into the predictive time horizon.

The joint modelling approach, like survival analysis, has its origins in the area of medical research. The commonly found applications are associated with the occurrence of some clinical events and repeated measurements of biomarkers to evaluate the efficacy of treatments. Yet, regardless of the research discipline, statistical principles and methods apply to any longitudinal follow-up study. The standard joint model assumes two sub-models, one for the survival process and the other for the longitudinal outcome, which in our case is the endogenous TVC. The link between the two sub-models comes from an assumed latent relationship modelled as subject-specific random effects. Both sub-models can be disentangled, considering conditional independence given the random effects (see Section 2.2).

The medical-related setup is analogous to credit analysis since we have credit events of interest and performance measurements over time. Nevertheless, relevant differences between medical and credit applications determine the modelling approach's appropriateness. Among these differences are the dataset size, the available information, the evolution of the TVCs, and the discrete nature of time.

The datasets observed in medical applications that use joint models are smaller than in credit analysis. Joint models are computationally expensive, so alternative estimation procedures are required if the goal is to scale the approach to larger datasets. Also, the standard joint model considers only one longitudinal outcome, but, as we mentioned above, we can find more than one in credit data, which also

increases the computational cost. Further, lenders have different sources of information that they can leverage in the modelling framework. An example is the location of real estate when mortgage loans are granted. These data can be used, for instance, to include spatial effects in the model. Similarly, if we have performance variables, we can imagine that some of these variables, such as the unpaid principal balance, are highly correlated to previous months' levels. Lastly, a common assumption in joint models for medical applications is to treat time as continuous. On the other hand, credit data, specifically instalment loans, are typically recorded monthly and intrinsically discrete. Consequently, a significant motivation also stems from the need to customise the joint modelling approach to credit-related applications.

1.3 Objectives

This thesis addresses the following question: Can we improve dynamic predictions in a survival analysis context with credit-related applications when potentially endogenous time-varying covariates exist? To shape the research strategy, we focus on studying a reasonably new approach in this area: the joint modelling of longitudinal and survival data. That allows us to handle endogeneity and offers a predictive survival framework that updates dynamically when new data are provided. Therefore, we define the research hypothesis as follows: The joint modelling approach is a viable, flexible and appropriate methodological framework to predict the time to an event of interest in credit-related applications when endogeneity over the TVCs exists and can lead us to better performance than survival approaches used in the literature.

To test the hypothesis, we focus on developing methodological and programming tools that enable us to represent and evaluate the mutual evolution of survival and longitudinal processes in the credit context. These developments contribute to the research community by providing new and more powerful ways to analyse data, build credit survival models with endogenous TVCs, and ultimately complement the toolbox of decision-makers and practitioners in the lending industry.

Moreover, as mentioned, the joint modelling approach is new to the credit literature, and as such, we conceive significant innovations to make this approach more suitable. In concrete, our specific objectives respond to the following four

extents:

Nature of survival time

- To formulate the joint models and their corresponding performing metrics assuming time as discrete.

Evolution of credit TVCs

- To include autoregressive terms in the longitudinal outcome and explore performance improvements.

Scalability

- To estimate the joint modelling approach in a more efficient way that allows us to scale it to sample sizes such as those seen in credit-related applications.
- To develop a scalable methodology for out-of-sample individual survival prediction.
- To develop a model comparison methodology integrated into the inference procedure without performing extensive post-estimation calculations.

Available information

- To estimate joint models considering more than one endogenous TVCs.
- To include spatial and spatio-temporal interactions in the joint modelling framework.

1.4 Contribution to knowledge

In this thesis, we present several contributions to the literature that are developed in Chapters 3, 4 and 5.

In Chapter 3, we make two main contributions. Firstly, to our knowledge, we present the first work exploring joint models for discrete survival data in credit risk applications. We argue that the discrete-time assumption is not only more appropriate for these data, but we also see computational benefits compared to its continuous-time counterpart. Second, we reinforce the modelling of subject

heterogeneity by including both random effects and autoregressive terms in the longitudinal outcome. This decision is motivated by the serial correlation found in our data and the potential implications these autoregressive terms could have for prediction performance.

In Chapter 4, we make four contributions. First, from a methodological perspective, we propose a multivariate joint model for longitudinal and survival data that can be framed using integrated nested Laplace approximations (INLA) (Rue et al., 2009). As detailed in Section 2.2.4, INLA is a fast and accurate deterministic algorithm for approximating Bayesian inference. This algorithm allows us to scale the approach to larger sample sizes than the ones seen in standard joint models and makes the approach viable for credit-related datasets. To illustrate the implementation and assess the recovery of true parameter values, we perform simulation studies.

Moreover, since we build the models for prediction purposes, the standard setting is to apply them to new individuals (out-of-sample) whose records are not used in the estimation procedure. In this setting, we are interested in individual survival predictions which require the marginalisation of subject-specific parameters. In the literature on joint models, these predictions are computed via empirical Bayes or simulation schemes (Rizopoulos, 2012). This step can be computationally expensive, mainly if applied to several new individuals. To address this issue, our second contribution is to propose a methodology for individual survival predictions using the Laplace method (Tierney and Kadane, 1986). This gives us more accurate approximations than the empirical Bayes approach and, unlike simulation methods, can be applied to several new individuals without significantly increasing the computational costs.

Our third and fourth contributions are from an empirical point of view. Specifically, we apply for the first time a multivariate joint model approach in the context of credit risk, in particular, to predict the probability of full prepayment in a portfolio of consumer loans. Although Hu and Zhou (2019) use joint models to predict mortgage loan prepayment events and show performance improvements compared to survival models, the authors consider only the univariate case and time as continuous. Finally, we show that these multivariate approaches result in better discrimination and calibration performance than the traditional survival models used in the literature (Thomas et al., 2017).

Finally, in Chapter 5, we make four contributions to the literature. First, from a methodological perspective, we propose a discrete-time joint model with a flexible baseline hazard that includes spatial and spatio-temporal interactions. As in Chapter 4, we frame the model within the INLA methodology to estimate it in a vast mortgage loan dataset. We use a dataset with a total of 2,559,056 observations, and as far as we are aware, this is the largest sample size for which a joint model of this type has been applied.

To compare model specifications, our third contribution in this chapter is to propose a new implementation of the *cross-validated Dynamic Conditional Likelihood* (cvDCL), a recently proposed cross-entropy estimate of the posterior predictive conditional density (see Section 5.2.3). Our implementation takes advantage of the quantities already computed by INLA in the model estimation, making it faster than the benchmark.

Lastly, we apply the proposed approach to predict full prepayment events in US mortgage loans. The empirical results show that including the spatial components can consistently improve the performance of the joint model. Yet, when spatio-temporal effects are included in addition to the spatial main effects, the performance improvements are less conclusive.

1.5 Thesis structure

The thesis consists of 6 chapters.

Chapter 2 presents the main background on the methodologies used throughout the thesis and the corresponding literature in the context of credit risks and joint models. To this end, we separate the chapter into two sections: survival analysis and joint models of longitudinal and survival data.

In Section 2.1, we start by describing survival analysis in general terms, introducing commonly used terminology and assumptions such as proportional hazards. We then present the survival approaches with TVCs used in the credit risk literature. Then, we formally illustrate the differences between endogenous and exogenous TVCs and the mathematical consequences that each type entails. Finally, we describe prediction techniques in survival models in the presence of TVCs.

Section 2.2 introduces the joint model approach, the more standard specifications used in the literature, and the advantages over survival models when dynamic prediction is essential. We then describe the estimation procedure under simulation-based schemes employed in Chapter 3. Next, we present the model from the perspective of latent Gaussian models, which is the family of models that the INLA methodology estimate. Finally, we describe how the INLA methodology works.

In Chapter 3, we present the joint model with autoregressive terms. We start by situating its relevance and contribution in the literature to give way to detailing the methodology, estimation, evaluation metrics and how individual survival predictions are carried out. Next, we present a simulation study with different configurations to support the inference procedure. We then apply the proposed model to estimate the time to default event on a portfolio of US mortgages. Finally, we conclude the chapter with the main findings.

Chapter 4 addresses the situation where more than one longitudinal outcome is present. We reformulate the approach to make it suitable for INLA (see Section 4.2). That makes it computationally efficient and applicable in the credit context. In addition, we present a methodology that is theoretically more accurate than relevant benchmarks for estimating individual survival predictions. As illustrated in Section 4.2.3, this methodology is based on Laplace's method but is independent of INLA and can be used with other estimation procedures. In Section 4.3, we explore the adequacy of model inference through a simulation study. Next, we build multivariate joint models to estimate the prepayment events in a German consumer loan portfolio (see Section 4.4). A discussion in Section 4.5 concludes the chapter.

In Chapter 5, we propose the joint model of longitudinal outcome and survival data with spatial effects and spatio-temporal interactions (see Section 5.2). This approach captures the survival effect due to the spatio-temporal correlation between events occurring within a short period and nearby locations. Section 5.2.2 develops its inference through the INLA approach. In Section 5.2.3, to compare different specifications, we present a new implementation of the cross-entropy estimate of the posterior predictive conditional density. Next, in Section 5.3, we apply the proposed joint model to predict the time to full prepayment of mortgage loans in the US. Section 5.4 discusses the main findings.

Finally, in Chapter 6, we summarise each chapter's results and findings. We also discuss the limitations of these approaches and examine different options to extend the proposed methods for future developments.

Chapter 2

Background

This chapter provides the background on survival analysis in the context of credit risk literature and the improvements that the joint models of longitudinal and survival data approach can bring in this respect. First, Section 2.1 describes survival models, their main assumptions, and how predictions are made when TVCs are provided. Then, Section 2.2 describes the joint model approach, the advantages that this approach confers over standard survival analysis when dynamic predictions are important and how we can estimate the joint models faster.

2.1 Survival Analysis

2.1.1 Introduction

As mentioned in Chapter 1, survival analysis, rather than predicting the outcome within a fixed time horizon, such as classification analysis, concerns about predicting when the event is potentially occurring. This framework also allows us to handle censored observations and facilitates the inclusion of time-varying covariates. Researchers have studied survival methods for years and have successfully applied them in different areas (see Kalbfleisch and Prentice, 2002; Collett, 2015, for a comprehensive description). In the context of credit lending, for example, since the first application introduced by Narain (1992), many authors have further developed this approach. Banasik et al. (1999) show how survival models can be as good as those obtained by traditional techniques. Stepanova and Thomas (2001, 2002) present how survival analysis can be used in the application and

behavioural models. Further, Bellotti and Crook (2009a) apply the approach to relate the prediction of default to time-varying covariates, particularly by including macroeconomic variables such as unemployment and interest rates.

More recently, Djeundje and Crook (2019a), using B-splines parametrisation, include time-varying coefficients in a survival model to predict credit card defaults, showing that it can improve predictive performance. Bhattacharya et al. (2019) build a Bayesian survival model for competing risks to study prepayment and default events in a US mortgage portfolio. Wang et al. (2020) use a discrete survival approach with TVCs to predict mortgage defaults and measure coefficient uncertainty in stressed scenarios. Similarly, Calabrese and Crook (2020) propose a survival model with a flexible parametric link and the inclusion of spatial contagion effect to improve the forecasting of mortgage defaults in the UK.

There are several other applications in the context of credit risk forecasting. These include forecasting the bankruptcy of corporates (Shumway, 2001; Duffie, 2005; Creal et al., 2014), estimating the expected loan profits at the time of application (Ma et al., 2010), modelling the duration of foreclosures in mortgage loans (Pennington-Cross, 2010), implementing a pricing model for a mortgage lender (McDonald et al., 2010), to name a few.

In the following sections, we formally introduce survival analysis, the commonly used assumptions and how the predictions are carried out under this approach when TVCs are provided.

2.1.2 Hazard and Survival functions

In survival analysis, a time to event variable T is defined for a subject as the time from a meaningful starting point, such as the origination of the credit, to the occurrence of the event of interest, such as default, early repayment, etc. Commonly in the follow-up of the subjects, there are censored cases where the event of interest did not occur throughout the study or the subject dropped out before the study ended (incomplete records). If the event is the default, for example, this can happen when the customer is still repaying the credit by the end of the study or closes the account before the end of the study. The censoring time in these cases is defined as the time elapsed from the starting point to the last available observation, and the time to event is known to be right-censored.

Two other types of censoring can also arise; left and interval censoring. Left censoring happens when it is known that the event occurred before some time t, but the exact moment is unknown. Interval censoring, on the other hand, appears when it is known that the event occurred within some interval $T \in (t_1, t_2)$. In this thesis, however, we focus on right censoring, and the reader is referred to Chapter 3 of Kalbfleisch and Prentice (2002) for a thorough discussion on censoring mechanisms.

The distribution function of $T > 0$ is represented by $F(t) = P(T \leq t)$, which is the probability that the event occurs in time equal to or before t. Similarly, the survival function is defined as $S(t) = P(T > t) = 1 - F(t)$ and describes the probability of the event occurring at a time later than t.

If T is continuous, it can also be described by its probability density function $f(t) = \mathrm{d}F(t)/\mathrm{d}t = -\mathrm{d}S(t)/\mathrm{d}t$ or, more commonly, by the hazard function $h(t)$ which is a probability rate function that measures how probable it is that a subject who has not evidenced the event before a time t, will do it at the next instant. Formally,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d} \ln S(t)}{\mathrm{d}t}. \tag{2.1}$$

The following relationship is also met,

$$S(t) = \exp\left(-\int_0^t h(u)\, \mathrm{d}u \right), \qquad \text{for } 0 < t < \infty.$$

If T is discrete, i.e., T takes values on $\{t_i | t_i < t_{i+1}, i = 1, 2, \ldots\}$, then its distribution can be described by a probability function $f(t_i) = P(T = t_i)$, $i = 1, 2, \ldots$, or by a hazard function as

$$h(t_i) = P(T = t_i | T \geq t_i). \tag{2.2}$$

Observe that in the discrete case, the hazard function is a probability rather than a probability rate as in the continuous case. The survival function $S(t) = P(T > t)$ follows

$$S(t) = \prod_{i | t_i \leq t} (1 - h(t_i)),$$

which comes from applying the multiplication law of probability and can be thought as if the subject survives longer than t, then it must survive each point $t_i \leq t$ with conditional probability $P(T \neq t_i | T \geq t_i) = 1 - h(t_i)$.

2.1.3 The Cox Regression Model

Both Equations 2.1 and 2.2 condition on the subject not having committed the event and having survived at t. We can also condition on covariates associated with the subject. The most popular regression method that does that was proposed in Cox (1972).

Assume the covariates for a subject is represented by a p-dimensional vector $\boldsymbol{z}^\intercal = (z_1, z_2, \ldots, z_p)$. If T is continuous, the Cox model assumes a parametric form for the hazard ratio as follows

$$h(t;\boldsymbol{z})/h(t;\boldsymbol{z}_0) = \exp(\boldsymbol{z}^\intercal\boldsymbol{\beta}), \tag{2.3}$$

where $h(t;\boldsymbol{z}_0)$, called baseline hazard function, is an unknown function for a baseline level of $\boldsymbol{z}_0$ (e.g., $\boldsymbol{z}_0 = \mathbf{0}$) and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients.

If T is discrete, then the Cox model can assume a parametric form for the ratio of the odds of the hazard as follows

$$\frac{h(t;\boldsymbol{z})/(1-h(t;\boldsymbol{z}))}{h(t;\boldsymbol{z}_0)/(1-h(t;\boldsymbol{z}_0))} = \exp(\boldsymbol{z}^\intercal\boldsymbol{\beta}). \tag{2.4}$$

It can be shown that Equation 2.4 reduces to 2.3 if time is continuous by noting that the discrete hazard in continuous form would be $h(t;\boldsymbol{z})\delta t$, with δt a time interval. Replacing this term in Equation 2.4 and taking the limit as δt tends to zero gives Equation 2.3 (see Cox, 1972; Stepanova and Thomas, 2002).

Because this method, in its continuous or discrete version, makes no assumption of the nature of the baseline hazard function itself but only on the effects of the covariates over the hazard, it is denoted as a semiparametric model.

In principle, the linear predictor $\boldsymbol{z}^\intercal\boldsymbol{\beta}$ can be replaced with a more general expression to include interactions, time-varying covariates, time-varying coefficients or any known function of the covariates. When the linear predictor is time-independent, however, the expression in Equation 2.3 is referred to as Cox proportional hazard model (Cox PH model). That comes from the fact that any increase in a covariate is multiplicative to the hazard, and the effect will be the same for all values of t. Similarly, Equation 2.4 is known as Cox proportional odds model when the linear predictor is time-independent.

Since the Cox PH model is widely used, some authors refer to it just as the Cox model and to the model that includes time-varying covariates (or effects of the

covariates that are time-varying) as the extended Cox model or the Andersen-Gill model (Kleinbaum and Klein, 2012; Rizopoulos, 2012), as it does not assume that the hazard ratio is constant in time. However, to avoid confusion, we use the term Cox model as in Kalbfleisch and Prentice (2002), i.e. to refer to the general model that does not assume proportionality. If we do assume proportionality, we will mention it explicitly. Moreover, throughout this work, we are interested in the version of this method that includes time-varying covariates (TVCs) in the linear predictor.

2.1.4 TVCs in Survival Credit Risk Models

Lenders collect customer data in a panel design, letting them incorporate time-varying covariates (TVCs) into their risk models (Crook and Bellotti, 2010). The inclusion of TVCs allows for predicting the event dynamically as more information becomes available over time. By doing so, improvements in the accuracy of the predictions and a better understanding of borrowers' behaviour have been obtained (Stepanova and Thomas, 2002; Djeundje and Crook, 2018; Dirick et al., 2019; Calabrese and Crook, 2020). Two types of TVCs are commonly incorporated into the analyses: (1) the ones specific to the customer, such as spending and repayment amounts, the outstanding balance, arrears in instalments, among others, and (2) the ones that are not associated with a particular customer, such as macroeconomic time series (interest rate, unemployment rate, etc.).

In terms of survival credit risk models, both types of TVCs have been used in the literature. Stepanova and Thomas (2002), along with introducing a consistent procedure for coarse-classifying the covariates in the Cox model, show that including the interaction of time with the refinancing purpose of the loan gives a more interpretable result. Moreover, Stepanova and Thomas (2001) by including the monthly balance and the partial delinquency of the loan, show that the Cox model is as good as the traditional logistic regression model in terms of prediction performance, but also it enables one to estimate the profit from the loans over time. Bellotti and Crook (2009a) study how the economic condition over time affects the probability of default of credit card accounts by including macroeconomic variables in the Cox model.

We can mention other works that have included TVCs when modelling the time to event in the credit context. Leow and Crook (2014), for example, use repayment

amounts, credit limits and outstanding balances along with intensity models to estimate not just the default event but also the probability of transitioning to delinquency. In addition, Bellotti and Crook (2013) and Bellotti and Crook (2014) show how to stress test credit card portfolios using simulated economic scenarios, and Djeundje and Crook (2018) present a multi-stage delinquency model which incorporates repayment amounts, credit limits, macroeconomic variables and random effects into the survival function. More recently, Calabrese and Crook (2020) introduced a model to predict the time to default for UK mortgages with spatial effects and include TVCs such as the loan balance and the estimated property value, among others, in addition to macroeconomic variables.

However, when it comes to TVCs specific to the borrowers, there is an important distinction to make and few studies in the credit risk literature address it. TVCs specific to the borrower could be exogenous or endogenous to the survival status (see Kalbfleisch and Prentice, 2002, Ch. 6.3)[1]. Exogenous TVCs are those whose future paths are not affected by the event's occurrence but affect the outcome[2]. In contrast, endogenous TVCs are variables whose path is influenced by the survival status of the individual and therefore carry direct information on the time to the event. As discussed in Section 2.1.5, this distinction is important because it directly determines the appropriate approach.

2.1.5 Exogenous vs. Endogenous TVCs

Assume a sequence of observations $\{y_s\}_{s\leq t}$ coming from a generic TVC represented by $\{Y_s\}_{s\leq t}$, for $s = 1,\ldots,t$. Denote the survival time as T and represent it by a sequence of indicator variables $\{X_s\}_{s\leq t}$ such that $(x_1,\ldots,x_t) = (0,\ldots,0,1)$ if $T = t$, or $(x_1,\ldots,x_t) = (0,\ldots,0,0)$ if $T > t$ (censored). As mentioned above, if Y_s is exogenous its future path is not influenced by the survival status of the individual. Formally,

$$p(y_{\tilde{t}}|\{x_s,y_s\}_{s<\tilde{t}}) = p(y_{\tilde{t}}|\{y_s\}_{s<\tilde{t}}) \quad \text{for } \tilde{t} = 2,\ldots,t,$$

whereas if Y_s is endogenous, that does not hold since the survival status affects its future path. This distinction has implications for the parameter estimation, as seen below.

[1]The terms exogenous and endogenous are also known as external and internal, respectively (Rizopoulos, 2012).

[2]To this extend, note that all TVCs that are not specific to the borrower are exogenous.

The joint probability of both stochastic process $\{X_s, Y_s\}_{s \leq t}$ can be written as

$$p(\{x_s, y_s\}_{s \leq t}) = p(x_t, y_t | \{x_s, y_s\}_{s<t}) \times \\ \times p(x_{t-1}, y_{t-1} | \{x_s, y_s\}_{s<t-1}) \cdot \ldots \cdot p(x_1, y_1), \quad (2.5)$$

where any term on the right-hand side follows

$$p(x_{\tilde{t}}, y_{\tilde{t}} | \{x_s, y_s\}_{s<\tilde{t}}) = p(x_{\tilde{t}} | y_{\tilde{t}}, \{x_s, y_s\}_{s<\tilde{t}}) p(y_{\tilde{t}} | \{x_s, y_s\}_{s<\tilde{t}}) \quad \text{for } \tilde{t} = 2, \ldots, t. \quad (2.6)$$

In the exogenous case, the second term of the right-hand side of Equation 2.6 reduces to $p(y_{\tilde{t}} | \{y_s\}_{s<\tilde{t}})$, which does not depend on the survival process, hence Equation 2.5 follows

$$p(\{x_s, y_s\}_{s \leq t}) \propto p(x_t | y_t, \{x_s, y_s\}_{s<t}) p(x_{t-1} | y_{t-1}, \{x_s, y_s\}_{s<t-1}) \cdot \ldots \cdot p(x_1 | y_1),$$

where the joint probability is proportional to the standard multiplication of each hazard function at $\tilde{t}$ formed by the conditional probability $p(x_{\tilde{t}} | y_{\tilde{t}}, \{x_s, y_s\}_{s<\tilde{t}})$. For the endogenous case though, that does not hold.

The consequence of including a TVC into a Cox model is that, for any individual, the probability of surviving longer than time t given that we have measured the TVC until t is a survival function when the TVC is exogenous, meaning that the usual relationship between the hazard and survival functions holds. The estimation is obtained by maximising the Cox's partial likelihood (Cox, 1975). However, when the TVC is endogenous, the probability of surviving longer than t given that we have measured the TVC until t is equal to 1 (and is no longer a survival function) since we know that the individual is still "alive" at t and will, for sure, survive longer than t (see Kalbfleisch and Prentice, 2002, Ch. 6).

This mutual evolution between the survival data and the endogenous TVC directly affects how the prediction and estimation are made. We can no longer rely on the standard Cox procedure, which requires methodological alternatives. The joint model approach we introduce in Section 2.2 addresses this problem by assuming conditional independence between X_t and Y_t given an underlying random effect U.

Examples of exogenous TVCs in the credit modelling context are the macroeconomic variables such as the inflation rate, GDP, and unemployment rate (Bellotti and Crook, 2009a), where their paths may influence the rate of default over time.

Still, their future values are not affected by a loan's default. Some examples of the endogenous case are the spending and repayment amounts, outstanding balance and arrears in instalments.

2.1.6 Prediction Framework

As seen above, the distinction between the type of TVCs included in the model is relevant to the parameter estimation. Similarly, depending on the assumptions used when modelling the survival data, there are direct implications for how the predictions are made. Figure 2.1 illustrates the general setup faced when we want to predict the probability of the event occurrence in the following Δt periods, given that the subject has survived the previous t periods. When a TVC is included, we have also observed its records, represented by the star points in the figure. The question is how we use the collected data to make the prediction.

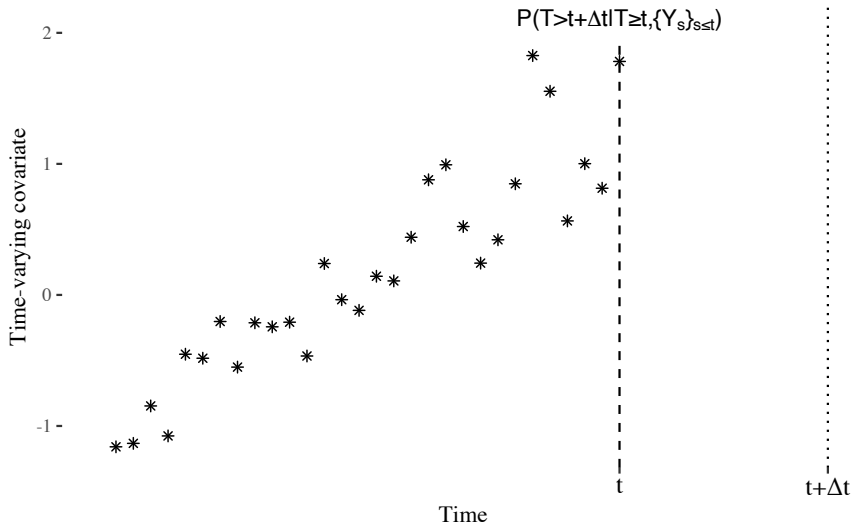

Figure 2.1: Historical measurements of a TVC until time t (left y-axis). The question is how to estimate the probability of surviving at $t + \Delta t$ given that we have observed this subject until time t.

One standard procedure in the literature is to lag the TVCs by the period we want to forecast. Figure 2.2 illustrates this procedure where the lagged values are the points coloured in blue, and on the right y-axis is the estimated survival probability with 95% credible intervals. In this way, if we want to predict the probability of the event in the next 12 months ($\Delta t = 12$), we estimate the model using the TVCs lagged in 12 months. Works that uses this procedure are, for

example, Bellotti and Crook (2009a); Leow and Crook (2016); Djeundje and Crook (2019a). However, this procedure has some drawbacks. First, it discards the data corresponding to the first lagged values. In the example of a time horizon of 12 months, the first 12 months are removed from the estimation. Moreover, by using the lagged values in the estimation, we do not relate the paths of the TVCs with the actual outcome. This desynchronisation might not be realistic when the TVC changes significantly over a period shorter than the forecast time horizon. In addition, since we need to define the time window of the prediction beforehand, it limits the analysis to other time windows. That goes against the flexibility of not depending on survival approaches' predefined time window. Finally, if the endogeneity between the TVCs and the survival process is significant, there is no guarantee that by lagging the TVCs, the possible bias is cleared.

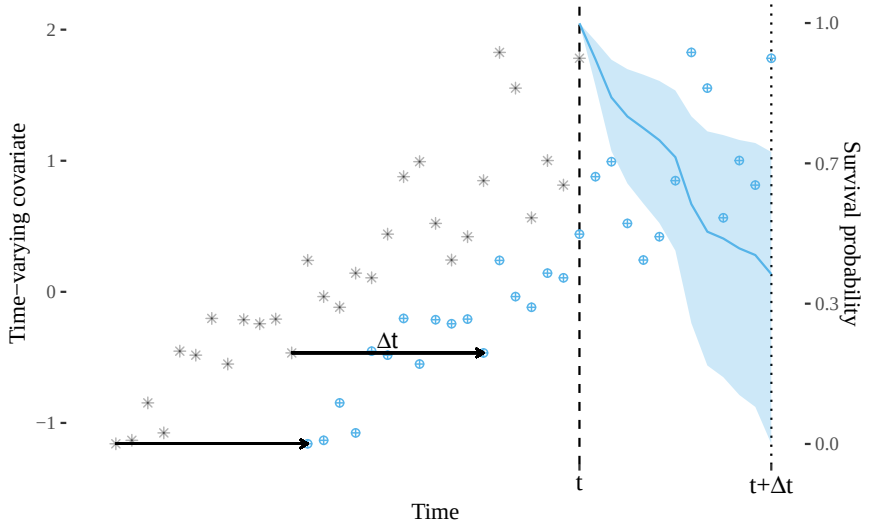

Figure 2.2: Estimation of the probability of surviving at $t + \Delta t$ given that we have observed this subject until time t following the lagged values' strategy.

Another alternative is to consider that our last record of the TVC is the most reliable one for future values. Hence, we carry this value throughout the prediction window (*Last Value Carried Forward*, Rizopoulos, 2012, Ch. 3). Figure 2.1 shows this procedure, where "future" values are coloured in orange, and the right y-axis depicts the survival probability. The advantage of this procedure concerning the one above is that we are not assuming any time window for the prediction in advance. Also, we do not need to discard information for the estimation. However, preserving the last value as fixed for future values is not sensible when the TVC changes considerably or we want to predict within a long-term horizon.

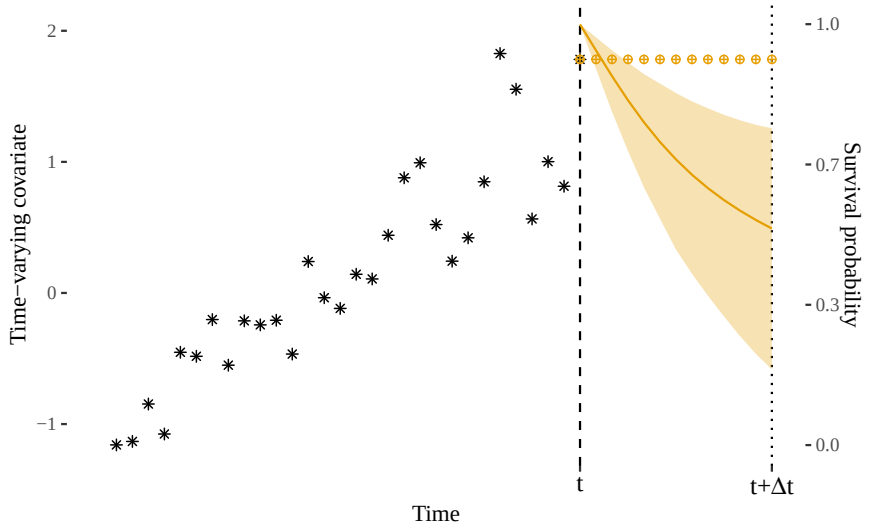

Figure 2.3: Estimation of the probability of surviving at $t + \Delta t$ given that we have observed this subject until time t following the last value carried forward strategy.

Finally, a joint model framework is a third approach and the one we are concerned about within this work. The basic idea is to learn the mutual evolution between the survival and the TVC processes to predict their future values. Figure 2.4 illustrates this procedure, where the solid green line estimates the expected value of the TVC and the dashed green line is the prediction. Again, the right y-axis shows the estimated survival probabilities. This procedure, detailed in the next section, can overcome the drawbacks of the last two methods. First, we are not limited to a specific time horizon to make the predictions. Moreover, we use all the available records and seize this information to make a prediction of the model-based TVC. Finally, unlike the other two methods, the joint model approach can handle the endogeneity among both processes.

2.2 Joint Models of Longitudinal and Survival Data

2.2.1 Introduction

The joint model of longitudinal and survival data approach (joint models hereafter) has been a rapidly evolving field of statistical methodology (Wu and Carroll, 1988; Tsiatis and Davidian, 2004; Henderson et al., 2000; Rizopoulos, 2012). Most of the literature on joint models comes from medical research where the interest

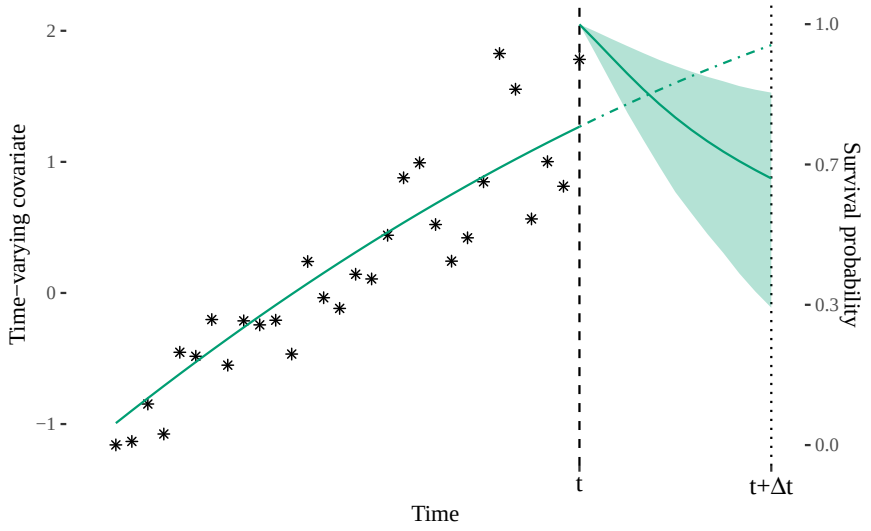

Figure 2.4: Estimation of the probability of surviving at $t + \Delta t$ given that we have observed this borrower until time t following the joint modelling strategy.

lies in the association between the repeated measurements of a biomarker for a patient and her survival time (Tsiatis et al., 1995). Still, the approach can be applied in any area where the association between both processes is of interest.

This approach addresses the endogeneity problem by modelling both the time to the event and the endogenous TVCs[3], simultaneously. In addition to avoiding estimation biases by considering the mutual evolution of both processes (Wulfsohn and Tsiatis, 1997; Tsiatis et al., 1995), it allows us to innately update survival probabilities when new observations of the TVCs are collected (dynamic prediction). Research in the medical context also shows that joint models increase the accuracy of the derived predictions (Rizopoulos et al., 2014).

The standard joint model is formed by two sub-models, one for the survival data and the other for the longitudinal outcome. Both are assumed to be conditionally independent given a latent structure. The survival process is commonly modelled by assuming a Cox model and a linear mixed-effects model for the longitudinal part (Rizopoulos, 2012). The two sub-models are associated through a functional form that could adopt many structures (see Hickey et al., 2016, for a detailed discussion about different structures). A thorough review of this modelling approach can be found in Tsiatis and Davidian (2004), who clarifies the main assumptions employed in the likelihood function. The textbooks, Rizopoulos (2012)

[3]To unify the jargon between the literature of joint models and credit risk, the endogenous TVCs will also be termed here as longitudinal outcomes.

and Elashoff et al. (2016), provide a comprehensive description of the technique, its inference and possible extensions. In particular, the former contributes with explicit applications programmed in the R software (R Core Team, 2021). Three other textbooks, Fitzmaurice et al. (2008); Wu (2009); Ibrahim et al. (2001), have specific chapters devoted to joint models. Moreover, Alsefri et al. (2020) summarises recent developments and issues.

In the credit-related context, to the best of our knowledge, there is only one paper that implements the joint model approach, Hu and Zhou (2019). Moreover, the authors provide promising predictive results compared to traditional survival models, strengthening the point for further studies. In Chapter 3, we offer a detailed comparison between our approach and theirs. One of the main differences, though, is that they assume time as continuous, while we consider it as discrete, providing us with some advantages as described in Section 3.1.

2.2.2 Discrete-Time Joint Model

As we mentioned in Section 1.2, assuming that the nature of time is discrete in a credit context is reasonable and commonly applied in related literature (Bellotti and Crook, 2014; Calabrese and Crook, 2020; Wang et al., 2020). However, in the literature on joint models, it is more typical to assume that survival time is continuous. In any case, some works study the discrete version, as we described below.

Albert and Shih (2010b) propose a two-stage approximation method for estimation in which the discrete hazard is modelled on the probit scale, which was extended later in Albert and Shih (2010a) to handle multiple longitudinal outcomes. Jaffa et al. (2011), more interested in the longitudinal process rather than the survival, introduce a joint model with bivariate longitudinal outcomes adjusted by informative right censoring. The authors then extended the approach for a high dimensional multivariate case (Jaffa et al., 2014). Furthermore, Barrett et al. (2015) propose an exact likelihood inference when the discrete hazard adopts a probit model by using distributional properties of the skew-normal family. They also include an unobserved stationary Gaussian process in the longitudinal model to bring more flexibility when the follow-up period is relatively long. Further, Bacci et al. (2018) assume a logit model for the discrete process and consider random intercepts in the longitudinal model to change over time according to an

autoregressive process of order 1.

Assume there are N subjects and the time to default for subject i ($i = 1, \ldots, N$) is represented by $T_i \in \mathbb{Z}_+$. We want to model T_i in terms of time-invariant covariates $\boldsymbol{z}_i$ and a longitudinal outcome $Y_{i,s}$ that is observed at times $s \in \{1, 2, \ldots, t_i\}$ where t_i is the time when either the event or the end of the follow-up happens. In theory, the number of observed values for the longitudinal outcome can differ from the survival times. However, throughout this thesis, we focus on equally spaced times and no missing observations before t_i (see Section 6.3 for a discussion in this regard). Analogously to the notation introduced in Section 2.1.5, we represent T_i as a sequence of binary indicators $X_{i,s}$ which is 1 if the event happens at time s and 0 otherwise. The standard assumption in the joint modelling approach is that $X_{i,s}$ and $Y_{i,s}$ are conditional independent given the random effects $\boldsymbol{U}_i$, i.e. $p(\{x_{i,s}, y_{i,s}\}_{s \leq t_i}) = \int p(\{x_{i,s}\}_{s \leq t_i} | \boldsymbol{U}_i) p(\{y_{i,s}\}_{s \leq t_i} | \boldsymbol{U}_i) p(\boldsymbol{U}_i) \, \mathrm{d}\boldsymbol{U}_i$ and interest is now turned on how to model each of the three elements of the integrand.

For the longitudinal part $p(\{y_{i,s}\}_{s \leq t_i} | \boldsymbol{U}_i)$, it is assumed that $Y_{i,s}$ can be described by an underlying signal $\eta_{Y\,i,s}$ and mutually independent noise terms $\epsilon_{i,s}$ as $Y_{i,s} = \eta_{Y\,i,s} + \epsilon_{i,s}$. Further, denote as $\boldsymbol{q}_{i,s}$ ($s = 1, 2, \ldots$) a vector of time-varying exogenous covariates (it could also be fixed in time). The term $\eta_{Y\,i,s}$ can be decomposed into fixed effects, $\boldsymbol{q}_{i,s}^\intercal \boldsymbol{\beta}_1$, and random effects, $\boldsymbol{d}_{i,s}^\intercal \boldsymbol{U}_i$, where $\boldsymbol{d}_{i,s}$ is the design vector at time s. This leads to the following mixed-effect model (Laird and Ware, 1982)

$$Y_{i,s} = \underbrace{\boldsymbol{q}_{i,s}^\intercal \boldsymbol{\beta}_1 + \boldsymbol{d}_{i,s}^\intercal \boldsymbol{U}_i}_{\eta_{Y\,i,s}} + \epsilon_{i,s}, \quad s = 1, \ldots, t_i, \tag{2.7}$$

where the subject-level random effects $\boldsymbol{U}_i$ are assumed as mutually independent and coming from a zero-mean multivariate Gaussian distribution of dimension r, $\boldsymbol{U}_i \sim N_r(0, \Sigma)$. The error terms are assumed normally distributed $\epsilon_{i,s} \sim N(0, \sigma^2)$, mutually independent and independent of $\boldsymbol{U}_i$.

For the survival part $p(\{x_{i,s}\}_{s \leq t_i} | \boldsymbol{U}_i)$ and following Allison (1982) discrete-time formulation, the probability that the event occurs at t_i is given by

$$p(\{x_{i,s}\}_{s \leq t_i} | \boldsymbol{U}_i) = \prod_{s=1}^{t_i} [p_{i,s}]^{x_{i,s}} [1 - p_{i,s}]^{1 - x_{i,s}}, \tag{2.8}$$

where $p_{i,s}$ is the conditional probability that subject i will commit the event at time s given both the random effects U_i and that it is "alive" at the beginning of s, i.e. $p_{i,s} = P(X_{i,s} = 1 | T_i \geq s, \boldsymbol{U}_i)$ (this is analogous to the discrete hazard

described in Section 2.1.2). Assuming, for instance, a logit link function as in Cox (1972), we can include the covariates in the following way

$$p_{i,s} = \text{logit}^{-1}(\underbrace{\nu_s + \boldsymbol{z}_i^\intercal \boldsymbol{\beta}_2 + \lambda f(\boldsymbol{U}_i, s)}_{\eta_{Xi,s}}), \tag{2.9}$$

where ν_s represents the baseline event time distribution. This term has been modelled in different ways, for example, as a set of constants for $s = 1, 2, \ldots$ (Allison, 1982), by cubic B-spline functions (Tutz and Schmid, 2016), or by random walk models, to mention a few (see Chapters 3, 4 and 5). Moreover, $\boldsymbol{\beta}_2$ is the vector of coefficients for the covariates $\boldsymbol{z}_i$ and λ is the association parameter between the survival and longitudinal processes. The function f relates both processes through the random effects $\boldsymbol{U}_i$ and, eventually, the time s. As mentioned before, f can adopt different structures (Hickey et al., 2016) but the standard one corresponds to the linear predictor, i.e. $f(\boldsymbol{U}_i, s) = \eta_{Yi,s}$ (Rizopoulos, 2012).

2.2.3 Estimation via MCMC

Following the notation used so far, assume the observed survival data for subject i ($i = 1, \ldots, N$) is denoted as $\boldsymbol{x}_i = \{x_{i,s} : s = 1, \ldots, t_i\}$ and the longitudinal measurements as $\boldsymbol{y}_i = \{y_{i,s} : s = 1, \ldots, t_i\}$, and represent the complete set of observations by $\mathcal{D} = \{\boldsymbol{y}_i, \boldsymbol{x}_i : i = 1, \ldots, N\}$[4]. The parameters to be estimated are those associated with the fixed effects $\boldsymbol{\beta}_1$, the covariance matrix of the random effects Σ, the variance of the error terms σ^2, the discrete baseline ν_s and its eventual hyperparameters $\boldsymbol{\tau}_\nu$ (depending on the modelling approach followed), the covariate coefficients $\boldsymbol{\beta}_2$ and the association parameter λ. Denote the set of all these parameters as Θ, thus the posterior distribution of Θ given $\mathcal{D}$ follows

$$\begin{aligned} p(\Theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\Theta)p(\Theta)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|\Theta)p(\Theta), \end{aligned} \tag{2.10}$$

where $p(\mathcal{D})$ is the marginal distribution of the data, $p(\Theta)$ is the prior distribution and $p(\mathcal{D}|\Theta)$ is the observation density, or likelihood (if the data are regarded as fixed), of the joint model. For consistency with the literature, denote this last

[4]Note that all the previously mentioned covariates are also observed, but we intentionally omit them to avoid notation overload.

term as $\mathcal{L}(\Theta|\mathcal{D})$ which can be decomposed as

$$\begin{aligned}\mathcal{L}(\Theta|\mathcal{D}) &= \prod_{i=1}^{N} \int p(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{U}_i, \Theta) p(\boldsymbol{U}_i|\Theta) \mathrm{d}\boldsymbol{U}_i \\ &= \prod_{i=1}^{N} \int p(\boldsymbol{x}_i|\boldsymbol{U}_i, \Theta) p(\boldsymbol{y}_i|\boldsymbol{U}_i, \Theta) p(\boldsymbol{U}_i|\Theta) \mathrm{d}\boldsymbol{U}_i,\end{aligned} \tag{2.11}$$

where we have used the assumption of conditional independence between the survival and longitudinal processes.

Moreover, by the Gaussian assumption on the random effects $\boldsymbol{U}_i$, the last term of the integrand in Equation 2.11 follows

$$\begin{aligned}p(\boldsymbol{U}_i|\Theta) &= p(\boldsymbol{U}_i|\Sigma) \\ &= (2\pi)^{-r/2} \det(\Sigma)^{-1/2} \exp\left(-\boldsymbol{U}_i^{\intercal} \Sigma^{-1} \boldsymbol{U}_i / 2\right).\end{aligned} \tag{2.12}$$

Additionally, by the assumption that the error terms are independent and zero-mean Gaussian distributed, the second term of the integrand in Equation 2.11 can be expressed as

$$\begin{aligned}p(\boldsymbol{y}_i|\boldsymbol{U}_i, \Theta) &= \prod_{s=1}^{t_i} p(y_{i,s}|\boldsymbol{U}_i, \Theta) \\ &= \prod_{s=1}^{t_i} (2\pi\sigma^2)^{-1/2} \exp\left(-(y_{i,s} - \eta_{Yi,s})^2 / 2\sigma^2\right),\end{aligned} \tag{2.13}$$

where $\eta_{Yi,s}$ is the linear predictor described in Equation 2.7.

For the first term of the integrand in Equation 2.11 and following Equation 2.8 we write

$$p(\boldsymbol{x}_i|\boldsymbol{U}_i, \Theta) = \prod_{s=1}^{t_i} [p_{i,s}]^{x_{i,s}} [1 - p_{i,s}]^{1 - x_{i,s}}, \tag{2.14}$$

where $p_{i,s}$ follows Equation 2.9 and if $f(\boldsymbol{U}_i, s) = \eta_{Yi,s}$, then

$$p_{i,s} = \mathrm{logit}^{-1}\left(\nu_s + \boldsymbol{z}_i^{\intercal} \boldsymbol{\gamma} + \lambda \eta_{Yi,s}\right). \tag{2.15}$$

Equations 2.12, 2.13, 2.14 and 2.15 completely specify the observation density in Equation 2.11.

Conceptually, this model can be estimated by maximising the likelihood from Equation 2.11, or the log-likelihood. Algorithms such as the Expectation Maximisation, Newton's method or modifications of them with asymptotic approximations have been used in the literature (Rizopoulos, 2012). However, the Bayesian

approach has some practical advantages in this context. First, asymptotic approximations are not required since inference is based on the full posterior $P(\Theta, \boldsymbol{U}|\mathcal{D})$, where $\boldsymbol{U}$ is the total set of random effects. Moreover, the computational implementation does not need tailored procedures, for example, to compute standard deviations, thus providing more flexibility when analysing different models' specifications (see Ibrahim et al., 2001, for more details).

To perform the Bayesian inference, though, we need to define the prior distributions on the parameters $p(\Theta)$ (Equation 2.10). The specific ones used in this work are described where appropriate. Some commonly used priors in this context are noninformative uniform ones across each parameter's domain for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\tau}_\nu$, λ and σ^2. For the covariance matrix, Σ, one traditional prior is the inverse-Wishart distribution or less heavy-tailed ones as the LKJ distribution family (Lewandowski et al., 2009).

There are several statistical programming languages available for Bayesian inference. The user needs to specify the proposed model's data, likelihood, and priors and find efficient parametrisation to avoid mixing problems and accomplishing fast inference. Well-known languages are BUGS, JAGS, Pyro, TensorFlow Probability, PyMC3 and Stan, all publicly available. In Chapter 3, we implement six models in *Stan* with the No-U-Turn Sampler (Hoffman and Gelman, 2014), a faster extension to Hamiltonian Monte Carlo algorithm (HMC). In addition, we can further increase inference speed by parallel sampling between and within chains. Regardless, we realise that to scale the joint model approach to the multivariate case and sample sizes more in line with credit risk applications; we need to find an alternative estimation approach which is the topic described below, in Section 2.2.4.

2.2.4 Joint Model as Latent Gaussian Model

Simulation-based MCMC schemes are computationally expensive or even infeasible for some applications with large $\mathcal{D}$, which is often the situation seen in the credit risk context. An alternative is to use the so-called integrated nested Laplace approximation methodology (INLA, Rue et al., 2009). INLA is a deterministic algorithm that provides accurate estimations of the posterior at a lower computational cost and is easily accessible through the `R-INLA` software package for R (`https://www.r-inla.org/`). This methodology applies to models belonging to

the class of Latent Gaussian models (LGM), which is the case of joint models as shown below.

Denote $\boldsymbol{\mu} = (\boldsymbol{\eta}_Y, \boldsymbol{\eta}_X, \boldsymbol{U}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\nu})$ as the set of all the unobserved variables in the joint model, where $\boldsymbol{\eta}_Y$ and $\boldsymbol{\eta}_X$ corresponds to the complete set of linear predictors described in Equations 2.7 and 2.9, each of them with $\sum_i^N t_i$ elements. The rest of the elements are latent variables and therefore $\boldsymbol{\mu}$ is referred as a latent field. Specifically, we assume the coefficients $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim N(\mathbf{0}, \tau_{\boldsymbol{\beta}}^{-1}\boldsymbol{I})$ with $\tau_{\boldsymbol{\beta}}$ a precision parameter and $\boldsymbol{I}$ is the identity matrix of the corresponding dimension. Similarly, we can assume that $\boldsymbol{\nu}$ is proportional to the normal Kernel $p(\boldsymbol{\nu}|\tau_\nu) \propto \exp[-\tau_\nu(\boldsymbol{\nu}^\intercal R_\nu \boldsymbol{\nu})/2]$, with R_ν a defined structure matrix (Rue and Held, 2005). Examples of models that can be specified in this way are the autoregressive and random walk models (see Chapters 4 and 5). Moreover, from Equation 2.12, we know that $p(\boldsymbol{U}_i|\Sigma) \sim N(\mathbf{0}, \Sigma)$ and denote $Q_{\boldsymbol{U}} = \Sigma^{-1}$ the precision matrix of the random effects.

Hence, $\boldsymbol{\mu}$ is a latent field distributed as a zero-mean multivariate Gaussian with precision matrix $Q(\boldsymbol{\theta}_1)$, with $\boldsymbol{\theta}_1$ the corresponding set of hyperparameters. Although the dimension of the matrix $Q(\boldsymbol{\theta}_1)$ can be very large, INLA takes advantage in terms of computation given the sparsity of the matrix (Rue et al., 2009).

Furthermore, denote as $\boldsymbol{\theta}_2$ the set of hyperparameters that have direct impact on the observation density. Then, we reformulate the observation density from Equation 2.11 to the INLA notation. First, recall that $p(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{U}_i, \Theta) = p(\boldsymbol{x}_i|\boldsymbol{U}_i, \Theta) \cdot p(\boldsymbol{y}_i|\boldsymbol{U}_i, \Theta)$, then each corresponding term can be written as

$$p(\boldsymbol{x}_i|\boldsymbol{U}_i, \Theta) = p(\boldsymbol{x}_i|\boldsymbol{\eta}_{Xi}, \boldsymbol{\theta}_2) = \prod_{s=1}^{t_i} p(x_{i,s}|\eta_{Xi,s}, \boldsymbol{\theta}_2)$$
$$p(\boldsymbol{y}_i|\boldsymbol{U}_i, \Theta) = p(\boldsymbol{y}_i|\boldsymbol{\eta}_{Yi}, \boldsymbol{\theta}_2) = \prod_{s=1}^{t_i} p(y_{i,s}|\eta_{Yi,s}, \boldsymbol{\theta}_2).$$

Note that each element of the observed data $\mathcal{D}$, namely $x_{i,s}$ and $y_{i,s}$, is associated with one element of $\boldsymbol{\mu}$, in this case $\eta_{Xi,s}$ and $\eta_{Yi,s}$, respectively. We can decompose the observation density $p(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{U}_i, \Theta)$ as the product of the elements of the observed data conditional to one element of the latent field, i.e. $\prod_{\tilde{s}=1}^{2t_i} p(\mathcal{D}_{\tilde{s}}|\mu_{\tilde{s}}, \boldsymbol{\theta}_2)$, with $\tilde{s}$ encoded accordingly.

Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and assume for $\boldsymbol{\theta}$ a prior $p(\boldsymbol{\theta})$. Thus, we have rewritten the

model in the following form

$$\begin{aligned}(\boldsymbol{\theta}) &\sim p(\boldsymbol{\theta}) \\ (\boldsymbol{\mu}|\boldsymbol{\theta}) &\sim N(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\theta})^{-1}) \\ \eta_j &= \sum_i c_{ji}\mu_i \\ (\mathcal{D}_j|\boldsymbol{\mu}, \boldsymbol{\theta}) &\sim p(\mathcal{D}_j|\eta_j, \boldsymbol{\theta}).\end{aligned} \tag{2.16}$$

The model described in Equation 2.16 is the type of model we can implement in INLA. Therefore, we can use this fast inference algorithm to estimate the discrete joint model.

For the sake of completeness, note that the posterior follows

$$\begin{aligned}p(\boldsymbol{\mu}, \boldsymbol{\theta}|\mathcal{D}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{\mu}|\boldsymbol{\theta}) \prod_{i=1}^{N} p(\mathcal{D}_i|\mu_i, \boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\boldsymbol{\mu}^\intercal \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{\mu} + \sum_j \log\{p(\mathcal{D}_j|\mu_j, \boldsymbol{\theta})\}\right].\end{aligned}$$

Section 2.2.5 below details how the estimation is carried out.

2.2.5 Estimation with INLA

We are interested in the posterior marginals, $p(\mu_i|\mathcal{D})$ and $p(\theta_j|\mathcal{D})$, specified by

$$\begin{aligned}p(\mu_i|\mathcal{D}) &= \int p(\mu_i|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta} \\ p(\theta_j|\mathcal{D}) &= \int p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}_{-j}.\end{aligned} \tag{2.17}$$

The INLA methodology computes these marginals based on the Laplace approximation (Tierney and Kadane, 1986). For $p(\boldsymbol{\theta}|\mathcal{D})$ this follows

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \left.\frac{p(\boldsymbol{\mu}, \boldsymbol{\theta}, \mathcal{D})}{p(\boldsymbol{\mu}|\boldsymbol{\theta}, \mathcal{D})}\right|_{\boldsymbol{\mu}=\text{arbitrary}} \approx \left.\frac{p(\boldsymbol{\mu}, \boldsymbol{\theta}, \mathcal{D})}{\tilde{p}_G(\boldsymbol{\mu}|\boldsymbol{\theta}, \mathcal{D})}\right|_{\boldsymbol{\mu}=\boldsymbol{\mu}^*(\boldsymbol{\theta})} := \tilde{p}(\boldsymbol{\theta}|\mathcal{D}),$$

where $\tilde{p}_G(\boldsymbol{\mu}|\boldsymbol{\theta}, \mathcal{D})$ is the Gaussian approximation to the full conditional and $\boldsymbol{\mu}^*(\boldsymbol{\theta})$ its mode. A crucial step in the procedure is to further approximate the terms $p(\mu_i|\boldsymbol{\theta}, \mathcal{D})$ by using the Laplace approximation one more time as

$$\begin{aligned}p(\mu_i|\boldsymbol{\theta}, \mathcal{D}) \propto \left.\frac{p(\boldsymbol{\mu}, \boldsymbol{\theta}, \mathcal{D})}{p(\boldsymbol{\mu}_{-i}|\mu_i, \boldsymbol{\theta}, \mathcal{D})}\right|_{\boldsymbol{\mu}_{-i}=\text{arbitrary}} &\approx \left.\frac{p(\boldsymbol{\mu}, \boldsymbol{\theta}, \mathcal{D})}{\tilde{p}_G(\boldsymbol{\mu}_{-i}|\mu_i, \boldsymbol{\theta}, \mathcal{D})}\right|_{\boldsymbol{\mu}_{-i}=\boldsymbol{\mu}^*_{-i}(\mu_i, \boldsymbol{\theta})} \\ &:= \tilde{p}(\mu_i|\boldsymbol{\theta}, \mathcal{D}).\end{aligned}$$

The approximations of integrals in Equation 2.17 are performed by the following steps:

1. Explore the hyperparameters joint posterior $\tilde{p}(\boldsymbol{\theta}|\mathcal{D})$ in order to construct a numerical integration grid $\{\boldsymbol{\theta}_w, \Delta_w\}$ for $\boldsymbol{\theta}$, where Δ_w are integration weights. The normalisation constants are also estimated in this step.

2. Approximate the marginal posterior density $p(\theta_j|\mathcal{D})$ through an interpolation algorithm based on the constructed grid $\{\boldsymbol{\theta}_w, \Delta_w\}$ (see Martins et al. (2013) for further details on the interpolation).

3. Construct $\tilde{p}(\mu_i|\boldsymbol{\theta}_w, \mathcal{D})$ for each w.

4. Approximate the marginal posterior $p(\mu_i|\mathcal{D})$ by $\sum_w \tilde{p}(\mu_i|\boldsymbol{\theta}_w, \mathcal{D})\tilde{p}(\boldsymbol{\theta}|\mathcal{D})\Delta_w$.

This is the original INLA procedure, however other computational efficient modifications for $\tilde{p}(\mu_i|\boldsymbol{\theta}, \mathcal{D})$ are also implemented in the `R-INLA` package and detailed in Rue et al. (2009).

Chapter 3

Discrete-Time Joint Model with Autoregressive Terms

This chapter is based on the manuscript Medina-Olivares et al. (2022a).

The chapter is organised as follows. Section 3.1 contextualises the discrete-time joint model with autoregressive terms in the relevant literature and describes the main contributions. Section 3.2 details the methodology for the proposed model and how the inference and the individual survival predictions are performed and assessed. Section 3.3 presents a simulation study considering different settings to test the accuracy and computational load of the proposed inference procedure. Section 3.4 presents the empirical results of the proposed model applied to a US mortgage portfolio. The concluding remarks follow in Section 3.5.

3.1 Introduction

This chapter describes the discrete-time joint model with autoregressive terms, how to perform its inference and the application that motivates it. Specifically, we are interested in predicting the time to default in the presence of endogenous time-varying covariates for fixed-rate US mortgages. We use the Single Family Loan-Level Dataset from Freddie Mac that is publicly available. The default is the event associated with the inability of a borrower to pay promptly. Although the specific definition of default depends on the requirements of the local regulator and the risk-averse nature of the financial institution in question, the Basel capital

framework defines default as the moment at which the borrower is past due more than 90 days in any credit obligation (BCBS, 2004). This is the standard baseline definition used among practitioners and the one used here.

We make two contributions to the literature. First, to the best of our knowledge is the first time that joint models for *discrete survival data* are empirically tested in credit risk applications. When developing this project, no works in the credit risk literature adopted a similar framework, which was part of our motivation to pursue it, in addition to the success it had shown in medical research. A recent study, though, Hu and Zhou (2019), implemented for the first time a joint model for predicting defaults on a peer-to-peer dataset and early repayments on mortgage loans. The authors provide promising results which, from our perspective, reinforce the importance of further investigations.

While Hu and Zhou (2019)'s work has similarities with ours in considering a joint modelling framework, our proposal differs in several respects. From an application standpoint, we predict different credit events using the same mortgage dataset. The authors predict early repayments; in our case, we are interested in credit defaults. That has implications for selecting the relevant longitudinal outcome[1] and in analysing the results when, for example, a highly imbalanced class is present, as when considering the event of default (see further details in Section 3.4).

From a methodological point of view, Hu and Zhou (2019) apply a joint model assuming time as continuous, which is the common assumption in the joint model literature (Lawrence Gould et al., 2015; Hickey et al., 2016) and in the available software to fit them (Furgal et al., 2019). However, in credit risk analysis, we see at least three reasons why discrete-time survival analysis should be preferred over continuous-time. First, events are intrinsically discrete. As account records are observed monthly, the events are defined based on these discrete observations. For example, the event of default is defined as the time at which three payments have been missed. A payment is missed if it has not been made by the billing date, which is a specific day of the month. Therefore, in this case, the continuous-time approach approximates an intrinsically discrete phenomenon (Tutz and Schmid, 2016). Second, in the monthly observed data, there are, and it is to be expected

[1]The authors decided to predict the early repayment because they could not find a significant longitudinal outcome that relates to the default.

that there will be, many events in the same month. The continuous approach theoretically implies that there would not be, and various well-known methods are needed to "solve" this inconsistency. However, further consideration is not required with the discrete-time assumption because tied events are handled without problems. Finally, with the discrete-time approach, the probability estimations when TVCs are included require simple summations over time points rather than complex integrations (Bellotti and Crook, 2013). This makes the model less computationally costly and, as shown in the application presented in Section 3.4, it can scale to sample sizes with more than 285K observations (the dataset in Hu and Zhou (2019) has 85k observations).

The second contribution we make relates to the treatment of the longitudinal outcome. The standard way to handle subject heterogeneity in the literature on joint models is by including subject-specific random effects (see Section 3.2). We propose to reinforce this approach by also including autoregressive terms. This decision is motivated by our data's serial correlation and the possible implications these additional terms have for prediction performance. The serial correlation is measured by the empirical autocorrelation functions (ACF, see Pinheiro and Bates, 2006) for two standard random effects specifications, namely, *random intercept* and *random intercept and slope* (see Figure 3.1). In both specifications, we see serial correlation, which supports the inclusion of autoregressive terms in the longitudinal process. Furthermore, as we will see in the empirical results, out-of-sample predictions can also be improved.

It is important to stress that including the autoregressive terms in the longitudinal process is facilitated because the observations are equally spaced and indexed by a discrete variable (time). If we would like, for example, to have a continuous-time version of this model, we would first have had to generalise the discrete autoregressive process to a similar continuous one, formulating the problem via a high-order stochastic differential equation. Unfortunately, that is not so simple in practice and, to the best of our knowledge, we are unaware of works that have done it.

In total, we implement six models in the platform for statistical modelling *Stan*. The first is a discrete survival model, which is our benchmark. The other five are discrete joint models, all with different specifications for the longitudinal outcome to analyse the importance of the random effects and the autoregressive terms in

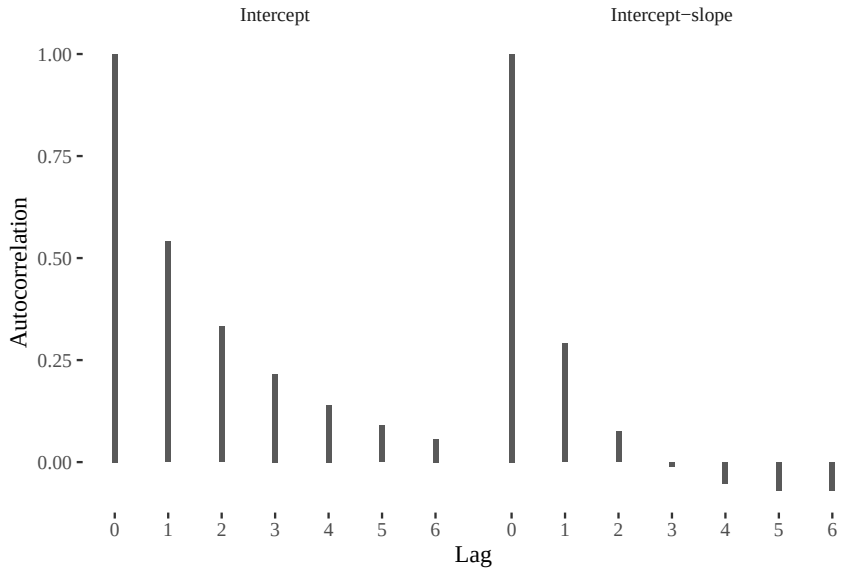

Figure 3.1: Empirical autocorrelation functions for the longitudinal outcome. On the left is the linear mixed-effect model with random intercept and, on the right, the linear mixed-effect model with random intercept and slope.

the predictions. Before applying the models to the mortgage dataset, in Section 3.3, we perform a simulation analysis to study how the most complex specification performs under different sample sizes. The results show the recovery of the actual parameter values and no signs of convergence problems.

Using discrimination and calibration metrics, we compare the six models by a five-fold cross-validation analysis for the mortgage dataset. The study exhibits two remarkable aspects. First, the discrete joint models can improve the discrimination compared to the traditional survival model. Second, this performance can be further enhanced when an autoregressive term is included in the longitudinal process, a difference that becomes even more pronounced as more historical data are considered.

3.2 Methodology

3.2.1 Joint model with autoregressive terms

We use the same notation introduced in Section 2.2.2 for the discrete joint models, i.e., assume that for borrower $i = 1, \ldots, N$ we are interested in modelling the time to default, denoted by $T_i \in \mathbb{Z}_+$, in terms of a vector of fixed covariates $\boldsymbol{z}_i$ and a longitudinal outcome $Y_{i,s}$ recorded at time points $s = 1, \ldots, t_i$. The last available

observation at time t_i is when either the default occurs or is right-censored. As before, represent T_i as a sequence of binary variables $X_{i,s}$ that takes the value of 1 if the borrower defaults at time s and 0 otherwise.

Note that the longitudinal process $Y_{i,s}$ from Section 2.2.2 is described by a predictor $\eta_{Yi,s}$ and independent error terms $\epsilon_{i,s}$ following $Y_{i,s} = \eta_{Yi,s} + \epsilon_{i,s}$, where $\eta_{Yi,s}$ is decomposed into fixed and random effects as $\eta_{Yi,s} = \boldsymbol{q}_{i,s}^\intercal \boldsymbol{\beta}_1 + \boldsymbol{d}_{i,s}^\intercal \boldsymbol{U}_i$, with $\boldsymbol{U}_i \sim N_r(0, \Sigma)$. To include autoregressive terms, suppose now that $Y_{i,s}$ is also described by an additional autoregressive structure of order p (see Hedeker and Gibbons, 2006, Ch.7), then the longitudinal process can be extended to

$$Y_{i,s} = \underbrace{\boldsymbol{q}_{i,s}^\intercal \boldsymbol{\beta}_1 + \boldsymbol{d}_{i,s}^\intercal \boldsymbol{U}_i + \sum_{r=1}^{p} \phi_r Y_{i,s-r}}_{\eta^*_{Yi,s}} + \epsilon_{i,s}, \quad s = p+1, \ldots, t_i, \tag{3.1}$$

where ϕ_r $(r = 1, \ldots, p)$ represents the coefficient for the r-th autoregressive term. Notice that with these additional terms, the longitudinal process is now correlated with its history and also with the subject-specific random effects $\boldsymbol{U}_i$. This might seem problematic from the inference perspective, however, as we described in Section 3.2.2 below, the conditional dependence structure for $Y_{i,s}$ given the random effects $\boldsymbol{U}_i$ follows an autoregressive structure with conditional expectation equal to the new predictor $\eta^*_{Yi,s}$ and conditional variance σ^2. This property allows a parallel representation of the observation density, which we use to speed up the MCMC sampling. In addition, to make the model well-specified, we assume that no event happened in the first p observations, so each borrower has at least p measurements.

For the survival process and following Equation 2.8, Section 2.2.2, we describe the observed sequence of $X_{i,s}$ $(s = p+1, \ldots, t_i)$, conditional on the random effects $\boldsymbol{U}_i$ and on the past observed values of Y_i, by[2]

$$p(\{x_{i,s}\}_{s \le t_i} | \boldsymbol{U}_i, \{y_{i,s}\}_{s < t_i}) = \prod_{s=p+1}^{t_i} [p_{i,s}]^{x_{i,s}} [1 - p_{i,s}]^{1 - x_{i,s}}. \tag{3.2}$$

The discrete hazard probability $p_{i,s} = P(X_{i,s} = 1 | \{X_{i,s^*} = 0\}_{s^* < s}, \boldsymbol{U}_i, \{y_{i,s^*}\}_{s^* < s})$ assumes a logit link function (Tutz and Schmid, 2016) as

$$p_{i,s} = \text{logit}^{-1}\Big(\underbrace{\nu_s + \boldsymbol{z}_i^\intercal \boldsymbol{\beta}_2 + \lambda_f f(\{y_{i,s^*}\}_{s^* < s}, \boldsymbol{U}_i, s)}_{\eta_{Xi,s}}\Big), \quad s = p+1, \ldots, t_i, \tag{3.3}$$

[2]We denote the realisations of a random variable in lowercase.

where ν_s is the baseline hazard function, $\boldsymbol{\beta}_2$ the coefficients for the time-fixed covariates $\boldsymbol{z}_i$, λ_f the association parameter between both processes and $f(\cdot)$ the function that relates them.

The terms ν_s are modelled with cubic B-spline functions (see Djeundje and Crook, 2018), this means that we represent $\nu_s = \boldsymbol{B}(s)^\intercal \boldsymbol{\alpha}$ with $\boldsymbol{B}(s)$ the vector of B-spline functions at time s and $\boldsymbol{\alpha}$ the corresponding vector of coefficients. Moreover, as noted before, the function f can adopt different structures (see Hickey et al., 2016) such as the ones used in this work and detailed in Section 3.4. Different structures of f will change the association parameter λ, so to avoid misleading comparison, we explicitly state the f dependency.

3.2.2 Estimation

Denote the complete observational data as $\mathcal{D} = \{\boldsymbol{y}_i, \boldsymbol{x}_i : i = 1, \ldots, N\}$ with $\boldsymbol{x}_i = \{x_{i,s} : s = p+1, \ldots, t_i\}$ and $\boldsymbol{y}_i = \{y_{i,s} : s = 1, \ldots, t_i\}$. Then, the complete set of parameters to estimate are $\Theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \lambda_f, \{\phi\}, \Sigma, \sigma^2\}$ and the observation density $\mathcal{L}(\Theta|\mathcal{D})$ is specified as

$$\begin{aligned} \mathcal{L}(\Theta|\mathcal{D}) &= \prod_{i=1}^{N} \int p(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{U}_i, \Theta) p(\boldsymbol{U}_i|\Theta) \mathrm{d}\boldsymbol{U}_i \\ &= \prod_{i=1}^{N} \int p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{U}_i, \Theta) p(\boldsymbol{y}_i|\boldsymbol{U}_i, \Theta) p(\boldsymbol{U}_i|\Theta) \mathrm{d}\boldsymbol{U}_i. \end{aligned} \tag{3.4}$$

Let us now describe the three terms in the integral of Equation 3.4. Note that the $\boldsymbol{U}_i$ follows a Gaussian distribution hence the last term follows

$$\begin{aligned} P(\boldsymbol{U}_i|\Theta) &= P(\boldsymbol{U}_i|\Sigma) \\ &= (2\pi)^{-r/2} \det(\Sigma)^{-1/2} \exp\left(-\boldsymbol{U}_i^\intercal \Sigma^{-1} \boldsymbol{U}_i / 2\right), \end{aligned} \tag{3.5}$$

where Σ is the covariance matrix of dimension $r \times r$.

The second term, $P(\boldsymbol{y}_i|\boldsymbol{U}_i, \Theta)$, can be decomposed by the chain rule over its

previous values in the following way

$$\begin{aligned} P(\boldsymbol{y}_i|\boldsymbol{U}_i,\Theta) &= P(y_{i,t_i},..,y_{i1}|\boldsymbol{U}_i,\Theta) \\ &= P(y_{i,t_i}|y_{i,t_i-1},\ldots,y_{i1},\boldsymbol{U}_i,\Theta)P(y_{i,t_i-1},\ldots,y_{i1}|\boldsymbol{U}_i,\Theta) \\ &= \prod_{s=1}^{t_i-p} P(y_{i,t_i-s+1}|y_{i,t_i-s},\ldots,y_{i1},\boldsymbol{U}_i,\Theta)P(y_{i,p},\ldots,y_{i1}|\boldsymbol{U}_i,\Theta) \\ &\propto \prod_{s=1}^{t_i-p} P(y_{i,t_i-s+1}|y_{i,t_i-s},\ldots,y_{i,t_i-s-p+1},\boldsymbol{U}_i,\Theta) \\ &= \prod_{s=1}^{t_i-p} (2\pi\sigma^2)^{-1/2}\exp\left(-(y_{i,t_i-s+1}-\eta^*_{Yi,t_i-s+1})^2/2\sigma^2\right), \end{aligned} \tag{3.6}$$

where we have used that the error terms are normally distributed with conditional variance σ^2 and that y_{i,t_i-s+1} $(s=1,\ldots,t_i-p)$ only depends on the previous p lags. Moreover, the terms $y_{i,p},\ldots,y_{i1}$ are not informative to the parameters and η^*_{Yi,t_i-s+1} is defined as in Equation 3.1.

The first term in the integral of Equation 3.4, $P(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{U}_i,\Theta)$, is conditionally dependent on the history of the longitudinal process $\boldsymbol{y}_i$. This dependency relates to the structure used for the link function $f(\{y_{i,s^*}\}_{s^*<s},\boldsymbol{U}_i,s)$ (see Equation 3.2). For the case of $f(\{y_{i,s^*}\}_{s^*<s},\boldsymbol{U}_i,s)=\boldsymbol{d}_{i,s}^\intercal\boldsymbol{U}_i$, for example, which is a standard structure in joint models (Hickey et al., 2016), we recover $P(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{U}_i,\Theta)=P(\boldsymbol{x}_i|\boldsymbol{U}_i,\Theta)$, as seen in Section 2.2. Yet, if $f(\{y_{i,s^*}\}_{s^*<s},\boldsymbol{U}_i,s)=\eta^*_{Yi,s}$ with $\eta^*_{Yi,s}$ following Equation 3.1, then we cannot separate $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ as before, since the event process now also depends on the previous p lag values of the longitudinal process. Let us assume the latter as a generalisation case of the other. Thus, following Equation 3.2 we write

$$P(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{U}_i,\Theta)=\prod_{s=p+1}^{t_i}[p_{i,s}]^{x_{i,s}}[1-p_{i,s}]^{1-x_{i,s}}, \tag{3.7}$$

with $p_{i,s}$ (see Equation 3.3)

$$p_{i,s}=\text{logit}^{-1}\left(\boldsymbol{B}(s)^\intercal\boldsymbol{\alpha}+\boldsymbol{z}_i^\intercal\boldsymbol{\beta}_2+\lambda_f\eta^*_{Yi,s}\right). \tag{3.8}$$

Hence, the observation density $\mathcal{L}(\Theta|\mathcal{D})$ (Equation 3.4) is fully specified by Equations 3.5, 3.6, 3.7 and 3.8.

The posterior distribution follows $P(\Theta|\mathcal{D})\propto\mathcal{L}(\Theta|\mathcal{D})P(\Theta)$ (Section 2.2.3). Thus, to complete the Bayesian model specification we need to define the prior distributions on the parameters, $P(\Theta)$. Specifically, we consider noninformative

uniform priors for the parameters λ_f, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\{\phi\}$ and σ, defined across each parameter's domain. Moreover, for the B-spline coefficients, $\boldsymbol{\alpha}$, we assume a multivariate Gaussian distribution $N(\mathbf{0}, \theta^2_{\boldsymbol{\alpha}} I)$, where $\theta_{\boldsymbol{\alpha}}$ is a hyperparameter with a *half-Cauchy* prior with a scale of 25. We found this parametrisation satisfactory to avoid mixing problems for the spline coefficients. However, to check the robustness of the results presented in Section 3.4, we perform in Appendix A.5 a robustness analysis using different priors. The study shows that the results are consistent.

Finally, a common choice for the covariance matrix, Σ, is to use a prior that belongs to the inverse Wishart's family. However, we experienced that this choice is not computationally efficient for the No-U-Turn Sampler (NUTS, Hoffman and Gelman, 2014), the sampling algorithm we use. Hence, following the suggestion in the platform's documentation for statistical modelling *Stan* (Stan Development Team and others, 2022), we decompose the covariance matrix as a correlation matrix and a vector of variances and define priors over these elements. For the correlation matrix, we use as a prior the LKJ distribution (Lewandowski et al., 2009). This distribution is specified by a regularisation parameter k, where a value of $k = 1$ represents a jointly uniform distribution over all possible correlation matrices. For values $k > 1$, the mode of the distribution is the identity matrix where the larger the k, the more sharply peaked is the distribution at the mode. In our case, we use a value of $k = 2$, which was found to work well for our simulation and application settings (see Appendix A.5 for further analysis). Furthermore, we set noninformative uniform priors in the positive domain for the vector of variances.

As mentioned before, we implement this and the other models specified in Section 3.4 in *Stan*. The sampling algorithm we use is NUTS, regarded as a faster extension of the Hamiltonian Monte Carlo algorithm (HMC).

3.2.3 Individual survival prediction

One of the advantages of the joint model approach is the dynamic prediction framework that it offers. Since joint models capture the mutual evolution of the survival and the longitudinal processes, we can exploit this learned relationship and extrapolate it into the future. For example, suppose we are interested in predicting, for a borrower, the probability of not defaulting in the next 12 months.

In that case, we can predict the path of the longitudinal outcome in the corresponding time window and get an estimation of the conditional probability given this future scenario. That differs from what is commonly done in the literature on credit risk, which is either to preserve the last observed value or to estimate the survival model with lagged TVCs, as described in Section 2.1.6.

Let us formalise now how the dynamic predictions are performed. Assume we are interested in predicting the default of a new borrower k not originally included in the data $\mathcal{D}$ used to estimate the joint model. Moreover, consider that this new borrower has not yet defaulted at least until time c and that we have collected, along with the fixed covariates, the longitudinal outcome up to that point (this commonly happens when loans are transferred between banks or when a customer opens a new account with an open-banking platform provider, where some of its historical data are shared). Denote these longitudinal records as $\boldsymbol{y}_k = \{y_{k,s} : s = 1, \ldots, c\}$. As noted before, our interest relies on estimating the probability of survival for a time window Δc in the future, given that we know the borrower has survived up to time c. In other words, we estimate the conditional probability of surviving time $c + \Delta c > c$ ($\Delta c \in \mathbb{Z}_+$) given that the borrower has survived up to time c. Mathematically, the expression follows $P(T_k > c + \Delta c | T_k > c, \boldsymbol{y}_k, \mathcal{D})$ but for readability purposes let us denote it as $\pi_k(c + \Delta c | c)$.

In order to estimate the expression $\pi_k(c + \Delta c | c)$, we need estimations of the random effects of borrower k, denoted as $\boldsymbol{U}_k$. However, since this borrower is not included in $\mathcal{D}$, Equation 3.5 does not provide estimations for it. One option to estimate $\boldsymbol{U}_k$ is to rerun the procedure described in Section 3.2.2, this time with a training set $\mathcal{D}$ that includes the historical data of the new borrower k, but this would be computationally expensive and not feasible if we apply it for many new borrowers coming at different times as is usually the case of credit-related applications.

A faster and more convenient option is to approximate $\pi_k(c + \Delta c | c)$ using empirical Bayes estimates for the random effects. This approximation procedure is detailed in Rizopoulos (2012) for the continuous-time setting. We follow it analogously for the discrete-time setting as explained below.

The conditional probability can be marginalised as

$$\pi_k(c + \Delta c | c) = \int P(T_k > c + \Delta c | T_k > c, \boldsymbol{y}_k, \Theta) P(\Theta | \mathcal{D}) \mathrm{d}\Theta, \tag{3.9}$$

where $P(\Theta|\mathcal{D})$ is the posterior distribution of the parameters given the sample $\mathcal{D}$, as described in Section 3.2.2 above. The first term of the integrand is the marginalisation over the random effects that reads

$$P(T_k > c+\Delta c|T_k > c, \boldsymbol{y}_k, \Theta) = \int P(T_k > c+\Delta c|T_k > c, \boldsymbol{y}_k, \boldsymbol{U}_k, \Theta) \times P(\boldsymbol{U}_k|T_k > c, \boldsymbol{y}_k, \Theta)\mathrm{d}\boldsymbol{U}_k. \tag{3.10}$$

Therefore, $\pi_k(c+\Delta c|c)$ is specified by both Equations 3.9 and 3.10. The approximation procedure is given by, first, approximating the integral of Equations 3.9 by choosing a representative posterior point-estimate $\hat{\Theta}$ from $P(\Theta|\mathcal{D})$. Then, the estimation of the random effects are obtained by solving $\hat{\boldsymbol{U}}_k = \text{argmax}_{\boldsymbol{U}}\{\log P(T_k > c, \boldsymbol{y}_k, \boldsymbol{U}|\hat{\Theta})\}$. Finally, the first order approximation is given by $\pi_k(c+\Delta c|c) \approx P(T_k > c+\Delta c|T_k > c, \boldsymbol{y}_k, \hat{\boldsymbol{U}}_k, \hat{\Theta})$, which in our discrete setting is given by

$$\begin{aligned}\hat{\pi}_k(c+\Delta c|c) &= \frac{P(T_k > c+\Delta c|\boldsymbol{y}_k, \hat{\boldsymbol{U}}_k, \hat{\Theta})}{P(T_k > c|\boldsymbol{y}_k, \hat{\boldsymbol{U}}_k, \hat{\Theta})} \\ &= \frac{\prod_{s=p+1}^{c+\Delta c}(1-\hat{p}_{k,s})}{\prod_{s=p+1}^{c}(1-\hat{p}_{k,s})} \\ &= \prod_{s=c+1}^{c+\Delta c}(1-\hat{p}_{k,s}),\end{aligned} \tag{3.11}$$

where $\hat{p}_{k,s}$ follows the specification described in Equation 3.8. In addition, to get standard error of Equation 3.11, we can estimate it through Monte Carlo simulation schemes as proposed in Rizopoulos (2011) and Proust-Lima and Taylor (2009).

3.2.4 Performance metrics

We evaluate the performance of the models presented in Section 3.4 under two dimensions: discrimination and calibration. Discrimination measures the model's ability to distinguish between borrowers that defaulted versus those who did not. However, the calibration measures how close or accurate the estimated probabilities are to what really happens.

There are different metrics to assess the model in terms of its discrimination and calibration capability. For discrimination, some common ones are the Area Under the ROC curve (AUC) (Fawcett, 2006), the Kolmogorov-Smirnov statistic (KS)

and the H-measure (Hand, 2009). For calibration, the Brier (Brier, 1950) and Logarithmic scores (Winkler, 1969)[3]. However, in their standard specification, these metrics do not control for right-censored cases and do not incorporate the new information collected once the evaluation time changes. Hence, we use generalisations presented in the joint model literature to address censoring and the dependence of the evaluation time. We adapt the notation for the discrete-time case as follows.

For discrimination, we use the AUC. This metric corresponds to the area enclosed by the curve formed by the proportion of correctly predicted events versus the proportion of incorrectly classified events considering all possible threshold values. The AUC ranges from 0.5 to 1, where a value of 0.5 represents a classifier that is not better than one that assigns labels randomly, and a value of 1 is a perfect classifier. Another interpretation of the AUC between evaluation times c and $c+\Delta c$ says that for any random pair of borrowers $\{i,j\}$ the metric can be described as (Hanley and McNeil, 1982)

$$AUC_c^{\Delta c} = P(\pi_i(c+\Delta c|c) < \pi_j(c+\Delta c|c)|\{T_i \in (c, c+\Delta c]\} \cap \{T_j > c+\Delta c\}),$$

where $\pi_i(c+\Delta c|c)$ follows Equation 3.9. Using this interpretation of the AUC facilitates the inclusion of censored cases as proposed by Rizopoulos et al. (2017). The way of doing it is to use model-based estimators of the censoring distribution by counting the concordant pairs of borrowers as the following

$$\widehat{AUC}_c^{\Delta c} = \widehat{AUC}_1(c,\Delta c) + \widehat{AUC}_2(c,\Delta c) + \widehat{AUC}_3(c,\Delta c) + \widehat{AUC}_4(c,\Delta c), \quad (3.12)$$

where each of these four AUC components is estimated over the four sets of possible combinations of concordant pairs, $\Omega_{ij}^{(l)}$, $l = 1,2,3,4$, defined respectively as

1. $\Omega_{ij}^{(1)}$: Borrower i experiences the default between times $c+1$ and $c+\Delta c$, and borrower j survives longer than $c+\Delta c$,
2. $\Omega_{ij}^{(2)}$: Borrower i is censored between times $c+1$ and $c+\Delta c$, and borrower j survives longer than $c+\Delta c$,
3. $\Omega_{ij}^{(3)}$: Borrower i experiences the event between times $c+1$ and $c+\Delta c$, and borrower j is censored between times $c+1$ and $c+\Delta c$,

[3]It is also common to show calibration plots, but as a visualisation aid rather than a formal metric.

4. $\Omega_{ij}^{(4)}$: Both borrower, i and j, are censored between times $c+1$ and $c+\Delta c$.

The estimates of the $\widehat{AUC}_l$, for $l = 1,2,3,4$, are specified as

$$\widehat{AUC}_l(c,\Delta c) = \frac{\sum_i^N \sum_{j\neq i}^N I(\hat{\pi}_i(c+\Delta c|c) < \hat{\pi}_j(c+\Delta c|c)) \cdot I(\Omega_{ij}^{(l)}) \cdot \hat{\nu}_{ij}^{(l)}}{\sum_i^N \sum_{j\neq i}^N I(\Omega_{ij}^{(l)}) \cdot \hat{\nu}_{ij}^{(l)}}.$$

$I(\cdot)$ is the standard indicator function, $\hat{\pi}_i$ is an estimate of the conditional probability (Equation 3.11) and the terms $\hat{\nu}_{ij}^{(l)}$ account for the probability that the pairs are comparable. Formally,

$$\begin{aligned}
\hat{\nu}_{ij}^{(1)} &= 1, \\
\hat{\nu}_{ij}^{(2)} &= 1 - \hat{\pi}_i(c+\Delta c|T_i), \\
\hat{\nu}_{ij}^{(3)} &= \hat{\pi}_j(c+\Delta c|T_j), \\
\hat{\nu}_{ij}^{(4)} &= (1 - \hat{\pi}_i(c+\Delta c|T_i))\hat{\pi}_j(c+\Delta c|T_j).
\end{aligned}$$

The calibration in the survival context is commonly evaluated by the expected error of predicting future events (Rizopoulos et al., 2017). Specifically, the general expression of the expected prediction error can be written as

$$EPE(c+\Delta c|c) = \mathbb{E}\Big[L\Big(N_i(c+\Delta c), \pi_i(c+\Delta c|c)\Big)\Big],$$

where the expectation is taken with respect to the distribution of event times. The expression $L(\cdot,\cdot)$ represents the loss function. This can be what we have mentioned above, i.e. the Brier score, and the Logarithmic score, among others. The expression $N_i(c+\Delta c) = I(T_i > c+\Delta c)$ indicates if the event did not happen before $c+\Delta c$. Since we are measuring an error metric, we expect that lower values mean better calibration performance.

To control for censoring in the framework above, we follow the proposal of Henderson et al. (2002) that is specified as

$$\widehat{EPE}(c+\Delta c|c) = \frac{1}{n(c)} \sum_{i:T_i>c} \Big[S_i(c+\Delta c|c) + E_i(c+\Delta c|c) + C_i(c+\Delta c|c)\Big], \quad (3.13)$$

where $n(c)$ represents the number of borrowers at risk at the beginning of time c and each of the terms in the sum are as follows

$$\begin{aligned}
S_i(c+\Delta c|c) &= I(T_i > c+\Delta c)L(1, \hat{\pi}_i(c+\Delta c|c)) \\
E_i(c+\Delta c|c) &= \delta_i I(T_i \leq c+\Delta c)L(0, \hat{\pi}_i(c+\Delta c|c)) \\
C_i(c+\Delta c|c) &= (1-\delta_i)I(T_i \leq c+\Delta c)\Big[\hat{\pi}_i(c+\Delta c|T_i)L(1, \hat{\pi}_i(c+\Delta c|c)) + \\
&\qquad + (1 - \hat{\pi}_i(c+\Delta c|T_i))L(0, \hat{\pi}_i(c+\Delta c|c))\Big],
\end{aligned}$$

where δ_i is the censor index of borrower i that takes the value of 1 if the borrower defaults and 0 otherwise.

As mentioned above, there are different options for choosing the loss function L. This work uses the Brier score (Brier, 1950) mainly due to its popularity. The Brier score corresponds to the mean squared error between the prediction and the true event status. Therefore, the estimate $\widehat{EPE}(c+\Delta c|c)$ presented in Equation 3.13 measures the mean square deviation at a time $c+\Delta c$ with historical data collected until c.

In addition to $\widehat{EPE}$, Henderson et al. (2002) also propose to measure the expected predicted error as an average over all the time points between $c+1$ and $c+\Delta c$. This metric nicely summarises the calibration for the whole interval in the following way

$$\widehat{PE}_c^{\Delta c} = \frac{\sum_{i:c<T_i\leq c+\Delta c}\delta_i w(c,T_i)\widehat{EPE}(T_i|c)}{\sum_{i:c<T_i\leq c+\Delta c}\delta_i w(c,T_i)}, \tag{3.14}$$

where $w(c,T_i) = \widehat{KM}(c+1)/\widehat{KM}(T_i)$ are weights to compensate for the loss of censored cases and $\widehat{KM}(\cdot)$ is the Kaplan-Meier estimator (Kaplan and Meier, 1958).

3.3 Simulation

We are interested in exploring how the MCMC scheme behaves under different sample sizes when we estimate the discrete joint model with autoregressive terms proposed in Section 3.2. For this, we generate three synthetic samples of sizes 1,000, 5,000 and 10,000 borrowers, respectively, over a maximum of 36 periods, where each period represents a different month. The default rates for these samples are 4.3%, 4.8% and 4.66%. The total number of borrower $\times$ time units are 24,424, 124,184 and 245,789, respectively.

For illustrative purposes, in Figure 3.2 we show the number of defaults that occurred in each duration time, considering the largest sample (10,000 borrowers). For instance, at time 10, 14 borrowers defaulted, and at time 30, 11.

In the same vein, Figure 3.3 shows, also for the largest sample, all the simulated paths of the longitudinal outcome. To facilitate the visualisation, we have highlighted ten borrowers who experience the default (dashed line in red) and ten who do not (dotted line in blue). For this setting, observe how the trajectories of

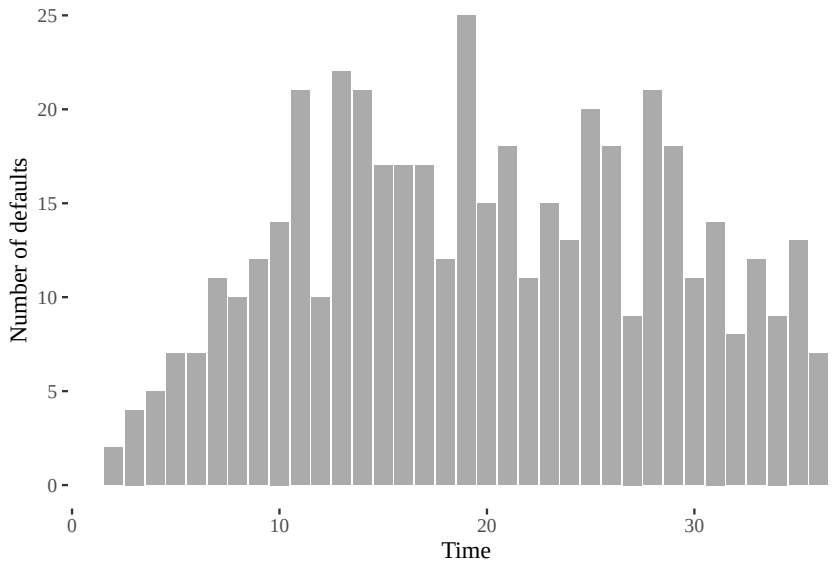

Figure 3.2: Distribution of events over time for simulated data with 10,000 subjects.

both types of borrowers are quite similar, especially at the beginning of the study. We see similar behaviour in our application because the setup for the simulation is precisely motivated by it (see Section 3.4). Although we will see that in the application we investigate different specifications of the joint model's link function f (see Equation 3.3), in this simulation study we focus on the link function $f(\{y_{i,s^*}\}_{s^*<s}, \boldsymbol{U}_i, s) = \eta^*_{Yi,s}$ with $\eta^*_{Yi,s}$ following Equation 3.1. This specification corresponds to the most general one.

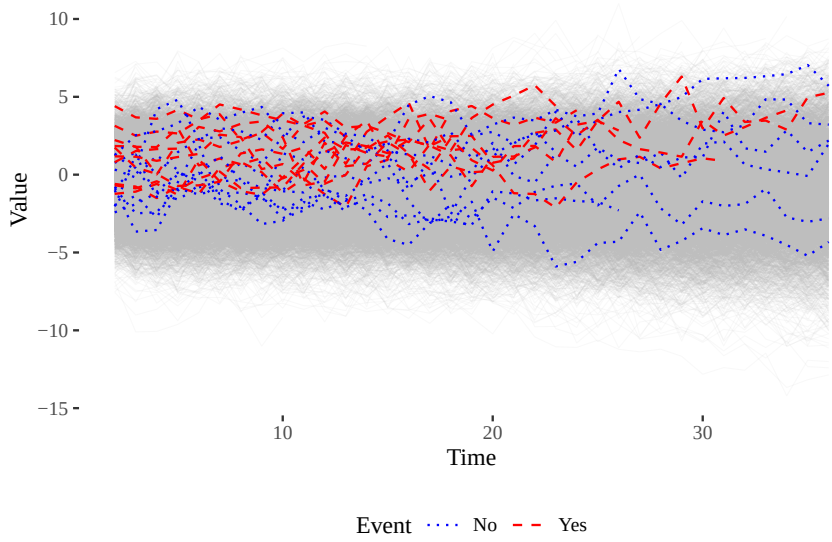

Figure 3.3: Simulated longitudinal outcome over time. Ten subjects that experience the event (dashed line) and ten that are censored (dotted line) are highlighted.

The specification of the generated longitudinal outcome $Y_{i,s}$ is represented by one fixed effect that plays the role of a general intercept and two random effects for each borrower, namely, the intercept U_{0i} and the slope U_{1i}. In addition, we include an autoregressive process of order one ($p = 1$), that is

$$Y_{i,s} = \underbrace{\beta_{01} + U_{0i} + U_{1i}s + \phi Y_{i,s-1}}_{\eta^*_{Yi,s}} + \epsilon_{i,s}$$

where we assume that $(U_{0i}, U_{1i})^\intercal \sim N_2(\mathbf{0}, \Sigma)$ and $\epsilon_{i,s} \sim N(0, \sigma^2)$. Moreover, we define the event process to depend on two covariates z_{1i} and z_{2i} that are fixed in time. Formally,

$$p_{i,s} = \text{logit}^{-1}(\nu_s + \beta_{12} z_{1i} + \beta_{22} z_{2i} + \lambda_f \eta^*_{Yi,s}).$$

The baseline hazard terms ν_s are simulated from a cubic polynomial function so that the overall default rate for the base reference is similar to those observed in mortgage loan portfolios. In this case, we consider approximately a 3% default rate over the 3-year horizon.

Once we have generated the data, we proceed to estimate the model via MCMC. We implemented the model in *Stan* with the No-U-Turn Sampler. We sample 3 independent chains with overdispersed starting points for each of the three simulation setups. Each chain has 4,000 and 2,000 iterations for the warm-up and sampling periods, respectively. Regarding the general inference diagnosis, none of the chains suffered from transitions that hit the maximum treedepth or were divergent. Furthermore, the energy Bayesian fraction of missing information (E-BFMI) was satisfactory for all transitions. In addition, all the estimated parameters had acceptable effective sample sizes $\hat{n}_{\text{eff}}$, which plays a similar role as the number of independent draws in the standard central limit theorem. Also, they all showed satisfactory potential scale reduction factors $\hat{R}$ which measures the consistency between chains by quantifying the between-chain over the within-chain variability. In summary, no problems were detected. Further details on the general diagnosis and these metrics are presented in Betancourt (2017).

The final 6,000 sampling iterations per simulation setup (2,000 per chain) are summarised in Table 3.1 by their means and 5%-95% posterior credible intervals in addition to the true generating parameter values.

Even though the baseline hazard ν_s is generated from a cubic polynomial function, we estimate it through cubic B-spline functions as described in Section 3.2.2. We

		$N = 1{,}000$			$N = 5{,}000$			$N = 10{,}000$		
	True	Mean	5%	95%	Mean	5%	95%	Mean	5%	95%
β_{12}	2.00	2.100	1.724	2.502	2.033	1.873	2.190	1.999	1.886	2.114
β_{22}	1.00	1.056	0.733	1.389	0.984	0.859	1.110	0.941	0.854	1.028
λ_f	1.00	1.027	0.862	1.204	1.002	0.932	1.075	0.997	0.944	1.048
β_{01}	-0.30	-0.320	-0.387	-0.255	-0.292	-0.318	-0.264	-0.289	-0.308	-0.270
ϕ	0.40	0.407	0.396	0.418	0.412	0.407	0.417	0.414	0.411	0.418
σ	1.00	1.001	0.993	1.009	1.000	0.996	1.003	1.003	1.000	1.005
σ_{U_0}	1.20	1.196	1.141	1.252	1.170	1.146	1.194	1.154	1.138	1.171
σ_{U_1}	0.05	0.049	0.046	0.052	0.048	0.047	0.049	0.049	0.048	0.050
ρ_U	-0.20	-0.182	-0.246	-0.117	-0.179	-0.209	-0.150	-0.184	-0.205	-0.163

Table 3.1: Estimations of the joint model with an autoregressive term over the different simulated samples.

use three internal knots at the 25th, 50th and 75th percentiles of the distribution of the event times (see Figure 3.2). That implies 7 spline coefficients to estimate since the degree of the functions is 3, we use 3 knots plus 1 coefficient due to the intercept, i.e. $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_6)^\intercal$ in Equation 3.8. We also explored other configurations with different numbers of knots, but no major improvements were obtained.

To illustrate how the baseline hazard is recovered for the whole time window, Figure 3.4 shows the simulated baseline hazard in a solid black line and, for each time point, the estimated 5-95% posterior credible intervals for the three settings. It is worth mentioning that all effective sample sizes $\hat{n}_{\text{eff}}$ of the αs are above 6000. Note that the intervals of the three settings cover the true value, and when more data are added, these are narrower, as expected. These results are in line with other works that have used B-spline functions to specify the baseline hazard (e.g. Djeundje and Crook, 2018; Bremhorst and Lambert, 2016).

Finally, and since the data generation process is known, one interesting question to explore in this simulation study is to measure how significant the bias in the parameter estimates is when we estimate a discrete survival model with the longitudinal outcome included as it is observed. This specification is relevant because it is a common practice in the credit risk literature when TVCs are present. The results are shown in Appendix A.1. However, in the empirical analysis conducted in Section 3.4, we cannot quantify the bias because we do not

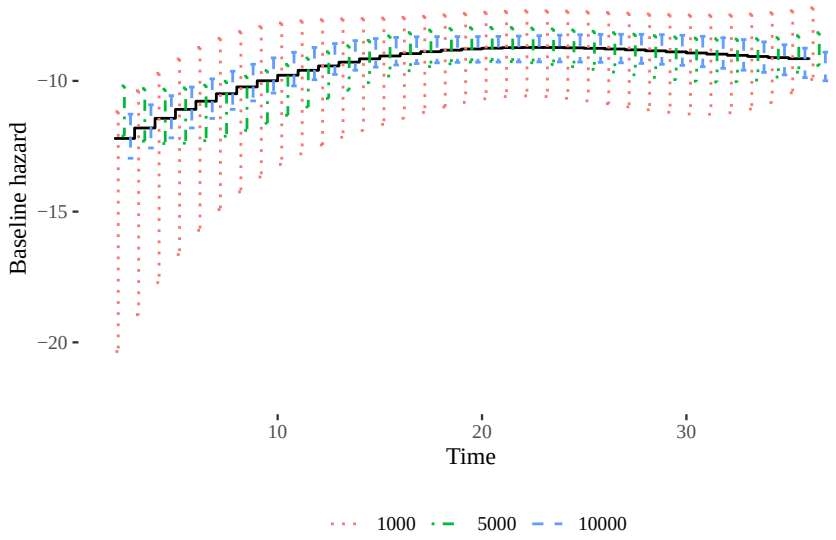

Figure 3.4: True baseline hazard ν_s (solid line) and the corresponding estimations for the three sample size settings with their 5-95% posterior credible intervals.

know the actual data generation process. Still, we can compare the predictions of each model.

3.4 Prediction of credit default in US mortgage portfolio

3.4.1 Data

We are interested in predicting credit default for US fixed-rate mortgages. The data provider is Freddie Mac, and the dataset we used is the Single Family Loan-Level Dataset which is publicly available[4]. This dataset contains loan-level granularity with application covariates and monthly performance records. Freddie Mac, since 1999, has been updating this dataset regularly, and for each vintage year, they also provide a randomly selected sample of 50,000 loans which is the one we use. In particular, we choose those that originated between October to December 1999 and follow their performance for the next 36 months.

The final number of loans in our training sample is 10,399 that corresponds to 285,462 observations. At the time of this writing, to the best of our knowledge, this is the largest sample size used in the literature on joint models. We use

[4] `https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset`

the common definition of default as the event when the borrower is 90 or more days past due. The percentage of loans that experience the default corresponds to 2.3% in the analysis period, and Figure 3.5 shows how they are distributed in time.

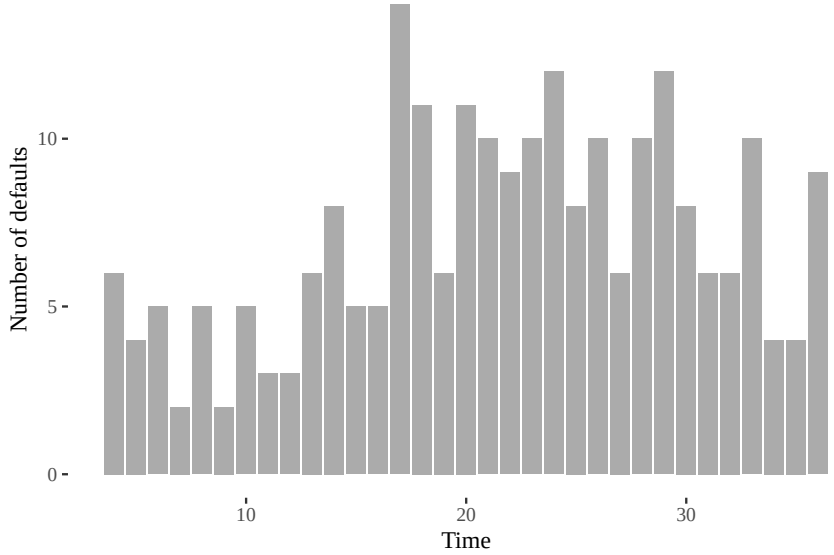

Figure 3.5: Distribution of the defaults over time for the training sample.

One of the challenges we faced in modelling default was to find a longitudinal outcome that was statistically significant. This problem is also mentioned in Hu and Zhou (2019) who, like us, use the Freddie Mac dataset and, not finding a variable to predict default, decide instead to focus on prepayment. In our case, we create a longitudinal outcome that nicely balances the scheduled repayments versus the actual repayments. We use the difference between the implicit interest rate and the fixed interest rate granted at origination, as described in the following.

Since we are provided with the loan amount, denoted as P_0, the fixed interest rate and the loan term, we can then calculate the original instalment amount A. With this information, in addition to the observed unpaid principal balance P_t, we can calculate an implicit interest rate i as shown in Equation 3.15. One direct result of the implicit interest rate is that if the payments are made as scheduled, then the implicit and the fixed interest rates are the same for all the periods. Otherwise, if there exists any unscheduled flow, it will then be reflected in the implicit interest rate.

$$P_t = P_0(1+i)^t - A\frac{(1+i)^t}{i} + \frac{A}{i}. \tag{3.15}$$

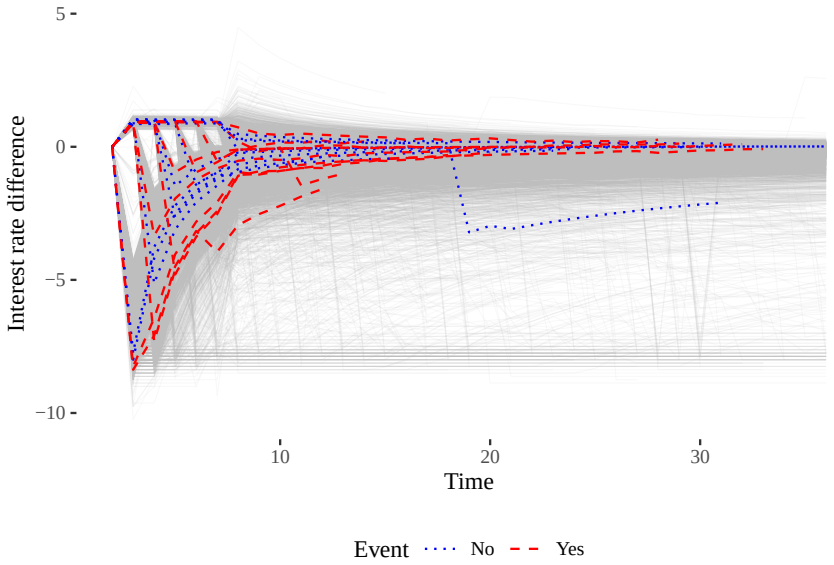

Figure 3.6: Evolution of the difference between the implicit and granted interest rate. Ten borrowers that defaulted (red dashed line) and ten who are censored (blue dotted line) are shown.

The final step to creating the longitudinal outcome is to take the difference between the implicit and fixed interest rates. The evolution of this variable for all the borrowers is shown in Figure 3.6. In that figure and for illustrative purposes, we highlight ten borrowers who default in dashed red lines and ten who do not in dotted blue lines. Note how the series either goes up or down in the first six months. This happens because the data provider reports, for the first six months, the current unpaid principal balance to the nearest $1000, which is consequently transmitted in the calculation of the implicit rate (if the rounded number is above or below the scheduled).

We described the predictor of the survival process in Equation 3.8 as the sum of the contributions of the baseline hazard, the link function and the time-invariant covariates. For the latter, we use the application covariates described in the following. These covariates are also in line with other works that have used this dataset (see Wang et al., 2020; Hu and Zhou, 2019)

- **fico** is a number summarising the borrower's creditworthiness (credit score) developed by FICO. Generally, the number disclosed is the score known at the acquisition time and used to originate the mortgage.
- **cltv** is the loan-to-value ratio based on the original mortgage loan amount plus any other mortgage loan amount divided by the mortgaged purchase price of

the property.

- **orig_upb** is the original unpaid principal balance of the mortgage on the note date.
- **dti** is the debt to income ratio. It corresponds to the borrower's monthly debt payments divided by the total monthly income used to underwrite the loan.
- **n_borr** is the number of borrowers obligated to repay the mortgage. Either one borrower (= 0, 38% of the loans) or more than one (= 1, 62% of the loans).
- **loan_purpose** indicates whether the mortgage loan purpose is a refinancing (= 0, 26% of the loans) or a purchase (= 1, 74% of the loans).

The descriptives statistics for the numeric covariates are shown in Table 3.2. To facilitate the MCMC sampling in each training sample, we standardise these covariates to have a zero-mean and standard deviation of 1 before estimation.

Covariate	N	Mean	SD	$Q_{2.5\%}$	$Q_{25\%}$	$Q_{50\%}$	$Q_{75\%}$	$Q_{95\%}$
fico	10399	710.70	52.50	619.00	672.00	716.00	753.00	786.00
cltv	10399	78.05	15.42	46.00	72.00	80.00	90.00	95.00
orig_upb*	10399	122.35	53.49	48.00	80.00	115.00	155.00	228.00
dti	10399	33.79	10.50	16.00	27.00	34.00	41.00	50.00

***1,000 USD.**

Table 3.2: Descriptive statistics for numerical covariates.

We analyse the models by five-fold cross-validation to assess them in an out-of-sample scenario, and we create each fold in such a way to preserve the overall default rate, given that this rate is already low. The default rate and the number of loans for each five-fold are shown in Table 3.3. Moreover, in Figure 3.7 we show the corresponding Kaplan-Meier curves where we observe similar survival behaviour.

Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
N	DFR (%)	N	DFR (%)	N	DFR (%)	N	DFR (%)	N	DFR (%)
2036	2.50	2093	2.25	2092	2.15	2037	2.26	2141	2.15

Table 3.3: Number of loans (N) and default rate (DFR) per fold.

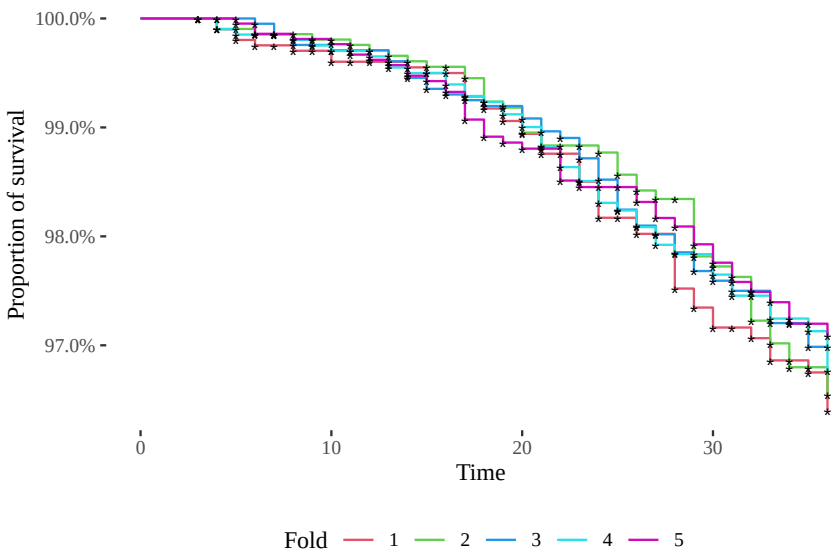

Figure 3.7: Kaplan-Meier curves per fold.

3.4.2 Models and results

We estimate six different models denoted as $M_0, \ldots, M_5$. All of them use the same application covariates in the survival process as described in Section 3.4.1. The differences come from the assumptions made on the link function f (Equation 3.3) and longitudinal outcome structure (Equation 3.1). These are summarised in Table 3.4.

Id	**Type**	**R-E**	**AR1**	$f(\cdot)$	$\eta^*_{Yi,s}$
M_0	Survival	-	-	$Y_{i,s}$	-
M_1	Joint	Int	No	U_{0i}	$\beta_{01} + U_{0i}$
M_2	Joint	Int	Yes	U_{0i}	$\beta_{01} + U_{0i} + \phi Y_{i,s-1}$
M_3	Joint	Int-slope	No	$\beta_{01} + U_{0i} + U_{1i}s$	$\beta_{01} + U_{0i} + U_{1i}s$
M_4	Joint	Int-slope	Yes	$\beta_{01} + U_{0i} + U_{1i}s$	$\beta_{01} + U_{0i} + U_{1i}s + \phi Y_{i,s-1}$
M_5	Joint	Int-slope	Yes	$\beta_{01} + U_{0i} + U_{1i}s + \phi Y_{i,s-1}$	$\beta_{01} + U_{0i} + U_{1i}s + \phi Y_{i,s-1}$

Table 3.4: Model specifications. **Id** is the model identifier, **Type** is survival or joint model, **R-E** specifies the random effects used (intercept only or intercept and slope), **AR1** if the model has autoregressive term. $f(\cdot)$ is the link function, however, for the survival model is the observed TVC (Equation 3.3). $\eta^*_{Yi,s}$ is the longitudinal predictor (Equation 3.1).

The first model M_0 corresponds to a discrete survival Cox model in which the longitudinal outcome, in our case, the interest rate difference, is included as observed. In other words, there is no joint modelling between the survival and

the longitudinal processes, and the latter is regarded as exogenous TVC. This specification is the standard assumption in credit risk literature when a survival model is considered (see, for example, Crook and Bellotti, 2010; Bellotti and Crook, 2013; Wang et al., 2020). As such, we treat M_0 as our benchmark. The other five models, M_1 to M_5, come from combining random intercept or random intercept and slope with or without the autoregressive term, as detailed in the table. This way, we better compare each component's importance in the joint model structure.

We perform five-fold cross-validation for each model specification with three independent chains for the MCMC sampling procedure. That is, we sample 90 posterior distributions in total since there are six specifications, each specification is estimated for each of the folds, and each fold comprises three independent chains. As in Section 3.3, the chains have a warm-up period of 4,000 iterations, and 2,000 sampling draws.

In computational terms, since each chain is independent of the other, we implement them in parallel (between-chain parallelisation). In addition, we increase the computational efficiency by adapting each likelihood specification to the recently released feature of the *CmdStan* interface for within-chain parallelisation[5]. Each chain run with 4 CPU cores of 16 GB of memory, and as a reference, one run of 6,000 samples for the most complex model (M_5) took 12 hours to finish. The Edinburgh Compute and Data Facility (ECDF, `http://www.ecdf.ed.ac.uk/`) provided the computational resources. Following the metrics mentioned in Section 3.3, no problems were detected concerning the general diagnosis of the NUTS sampler. That is, none of the transitions was divergent or hit the maximum treedepth, all had satisfactory E-BFMI as well as the $\hat{n}_{\text{eff}}$ and $\hat{R}$ for all the parameters.

The summary of the parameter estimates for the six models using the final 6,000 samples (2,000 per chain) is shown in Table 3.5. In this table, we present the results for one of the five folds (keeping fold one out in this case) since the others are consistent. We first note that their 5-95% posterior credible intervals for almost all parameters do not include 0. Second, there is strong evidence that the autoregressive coefficient ϕ for models M_2, M_4 and M_5 is significant. Furthermore, the posterior means of the parameters associated with the application

[5]See `https://mc-stan.org/users/interfaces/cmdstan.html`

covariates have somewhat similar estimates among the six models with agreeing signs. For example, higher values of *fico*, meaning better creditworthiness, are associated with a lower probability of default.

Moreover, the higher the loan to the purchase price (*cltv*), the higher the probability of default. A similar result is obtained for the debt to income ratio *dti*. Suppose more than one borrower is responsible for paying the loan (*n_borr*). In that case, we observe that the probability of default is also lower, and the same is noted when the loan purpose is to purchase the mortgage instead of refinancing it (*loan_purpose*). The exception comes from *orig_upb* estimated by model M_4 where its credible interval does include 0 and its estimated mean drops 50% in relation to the other models. In addition, the posterior samples of the association parameter λ_f show differences among the specifications as expected since the linking variables are not strictly comparable (for example, constant versus linear tendency), but all the intervals are far from 0. Nevertheless, the signs are all positive, which can be interpreted as if the level of the difference between the implicit and the original interest rate increases, then also the probability of default increases.

We measure the performance of the individual survival predictions under the discrimination and calibration metrics described in Section 3.2.4. Both the $\widehat{AUC}_c^{\Delta c}$ (Equation 3.12) and the $\widehat{PE}_c^{\Delta c}$ (Equation 3.14) depend on the evaluation time c and the forecast window Δc. We study the predictions for the range of $c \in [6, 24]$ and $\Delta c = 12$ to analyse how the models behave when more information is collected in time. For instance, if $c = 6$, we use the collected data until the sixth month and predict the probability of default for months 7 to 18. Further, all the predictions are made for the holdout fold, so the newly collected data are not used for estimating the parameters of the models but rather to estimate the random effects that serve the individual predictions as described in Section 3.2.3.

To compare the models against the benchmark (M_0), we calculate the difference in the $\widehat{AUC}_c^{\Delta c}$ for all the values of c within their corresponding fold. Table 3.6 shows the means and standard deviations of the difference in the AUC considering the five folds. It is worth noting that since we are calculating the standard deviation among the folds, we need to correct for overlapping training sets. To do so, we include the additional correlation term detailed in Nadeau and Bengio (2000). The first column of Table 3.6 corresponds to c, the number of months of known

data for the out-of-sample borrowers. We observe that for c between 6 and 9, models M_1 and M_2 outperform the benchmark in terms of discrimination for the forecast window of 12 months but for larger c, both remain practically the same to M_0. Moreover, for $c \leq 13$ there is not a great difference for models M_3, M_4 and M_5 with respect to M_0, however, for $c \geq 14$, the discrimination increases considerably, specially for M_4, with an average increase of more than 0.1 in the $\widehat{AUC}_c^{\Delta c}$.

In the same line, Table 3.7 shows the mean differences and standard deviations of $\widehat{PE}_c^{\Delta c}$ with respect to M_0, for the same range of c and forecast window. For ease of viewing, all the values are scaled by 100. For $c < 12$, we observe that the calibration metrics of the joint models are generally better than the benchmark, particularly for models M_4 and M_5. For $c \geq 12$, however, models M_3 and M_4 start to increase the expected predictive error in comparison to M_0. Model M_5 also increases the predictive error but not as much as models M_3 and M_4, which can be seen as a good balance between improvement in discrimination without affecting too much the calibration. Furthermore, models M_1 and M_2 recover the same performance levels as the benchmark.

Models M_3, M_4 and M_5 show better discrimination than the benchmark when more historical information is collected, but the same is not valid in calibration. This discrepancy stems mainly from the fact that these data are highly unbalanced, i.e. the number of defaults is considerably less than the number of non-defaults. Under these circumstances, it could happen that any model, for instance, that assigns a survival probability of 1 to all still has a relatively good calibration, so it is essential to take this metric with caution and understand where the major contributions come from.

Table A.3 in Appendix A.3 shows the 5-95% probability ranges estimated by each model and separated by non-defaulters (value 0) versus defaulters (value 1). We observe that, for $c > 12$, the joint models M_3, M_4 and M_5 start to have a broader range than the benchmark for both labels, which is also when the differences in the calibration metric appear. In other words, the joint models can identify better the defaulters versus the non-defaulters since they have better discrimination performance and assign lower probabilities of surviving to the defaulters than the benchmark. However, these models also assign lower probabilities to the non-defaulters, and because of the large number of these cases in the data, the

calibration is made worse.

To investigate to what extent the models are sensitive to class imbalance, we re-estimate the models M_0 (benchmark) and M_5 (joint model with the autoregressive term) by controlling the proportion of non-defaulters in the data. Appendix A.4 shows the results for two scenarios. The first randomly reduces the number of non-defaulters so that 75% of the loans are non-defaulters, and the second has an equal number of defaulters and non-defaulters. These results indicate that the joint model's relative calibration significantly improved compared to the Cox model. The difference between them ($\Delta\widehat{PE}_c^{12}$ M_5) for $c \geq 15$, is reduced in more than 50% when compared to the results shown in Table 3.7.

3.5 Discussion

The inclusion of TVCs into survival credit default models is widely applied in the literature to improve the predictions or enhance the understanding of why borrowers default (Bellotti and Crook, 2009a, 2014; Dirick et al., 2019; Wang et al., 2020; Calabrese and Crook, 2020). However, few works focus on distinguishing the type of variable included (see, for instance, Dirick et al., 2019; Hu and Zhou, 2019), thus treating endogenous and exogenous variables equally. This practice can lead to two main problems if the TVC is endogenous. First, from a statistical standpoint, we might encounter biased parameter estimations (Section 2.1.5 and Appendix A.1). Second, from a forecast perspective, we lack a dynamic prediction framework that takes advantage of the mutual evolution between the TVC and the survival time, forcing the prediction to keep the last observed value fixed or estimating the model with lagged values of the TVC (Crook and Bellotti, 2010; Bellotti and Crook, 2013; Wang et al., 2020).

To address the inclusion of endogenous TVCs into survival credit default models, we explore the joint modelling approach and adapt it to handle features typical of credit-related applications. First, to the best of our knowledge, this is the first work that uses discrete-time joint models in the credit context. Second, from a methodological angle, we propose an extended joint model that incorporates autoregressive terms into the longitudinal outcome. We take advantage of the fact that observations are equally-spaced and indexed by a discrete variable (time). This extension is motivated by the autoregressive components seen in the data

(see Figure 3.1) and how these additional terms could eventually improve the accuracy of the predictions.

In total, we implement six models, a standard discrete survival model (M_0) that is our benchmark and five joint models (M_1, M_2, M_3, M_4, M_5), all of them following the Bayesian paradigm, coded in *Stan* and using `CmdStan` interface with within-chain parallelisation feature (Stan Development Team, 2018). We study the most general case of the implementations (M_5) via simulation. The study shows a satisfactory converging diagnosis for three independent sampling chains and true value recovery for different sample sizes.

Furthermore, we apply all the models to US mortgage loan data and compare them via cross-validation analysis. The empirical results show that the joint models that assume the longitudinal outcome with only random intercepts, either with or without an autoregressive term (M_1 and M_2, respectively), only improve the discrimination measure compared to the benchmark when not much historical information of the new borrowers is known. Yet, the other three joint models, namely, M_3, M_4 and M_5, show a more remarkable improvement in terms of discrimination when more historical data are collected, especially M_4 that includes autoregressive correction in the longitudinal outcome.

In calibration, we see that when using the historical data up to the first year (12 months), the joint models are generally better than the benchmark. Moreover, when more historical data are considered, models M_1 and M_2 preserve the same calibration level as the benchmark. However, for models M_3, M_4 and M_5 the calibration error grows in comparative terms. That is because these models estimate posterior probability distributions with higher variability than the benchmark for the non-defaulters. Given that these data are highly imbalanced, greater variability in the probabilities is more detrimental to the overall quality of the calibration in comparison to the benchmark. Nevertheless, when we control by the class imbalance, we see that this difference is considerably reduced.

In this chapter, we include only one longitudinal outcome in the model with only one autoregressive term in the implementations. A potential extension of this work might be to consider a multivariate longitudinal case with different autoregressive orders, where more complex payment patterns can be recognised and included in the time-to-event prediction—for instance, being able to measure

and incorporate correlations between the use of the credit card and the implicit interest rate through a bivariate longitudinal model.

However, a significant drawback of this methodology is the computational cost when standard estimation procedures are employed. Financial institutions typically estimate models on big sample sizes, on thousands or millions of data. Suppose we want to explore more complex structures of joint models, such as one with multiple longitudinal processes. If we scale the approach for real-life applications, we might need to look for alternative estimation procedures faster than MCMC schemes. That is what we study in Chapter 4 where we reformulate the multivariate joint model approach to be estimated with the integrated nested Laplace approximation (INLA) (Rue et al., 2009), a fast deterministic algorithm for Bayesian inference.

To conclude, using joint models is a promising approach to credit-related applications in which we usually have a variety of endogenous TVCs that could bring relevant predictive information. Likewise, we believe our extension to include autoregressive terms in the longitudinal process can be further exploited to extract predictive behaviours and better understand the dynamic nature of credit defaults.

Parameter	M_0			M_1			M_2		
	Mean	5%	95%	Mean	5%	95%	Mean	5%	95%
fico	-0.701	-0.819	-0.584	-0.698	-0.813	-0.583	-0.697	-0.816	-0.576
cltv	0.515	0.334	0.705	0.544	0.361	0.732	0.542	0.367	0.728
orig_upb	-0.151	-0.294	-0.011	-0.183	-0.328	-0.037	-0.182	-0.323	-0.044
dti	0.152	0.025	0.284	0.165	0.033	0.292	0.166	0.041	0.294
n_borr	-0.260	-0.513	-0.007	-0.268	-0.521	-0.019	-0.264	-0.510	-0.020
loan_purpose	-0.977	-1.246	-0.697	-0.992	-1.268	-0.717	-0.987	-1.265	-0.713
λ_f	1.456	1.135	1.778	0.345	0.150	0.550	1.053	0.480	1.703
β_{01}				-0.472	-0.495	-0.450	-0.206	-0.216	-0.195
σ_{U_0}				1.209	1.193	1.226	0.504	0.496	0.512
σ				0.935	0.932	0.937	0.787	0.785	0.789
ϕ							0.587	0.584	0.591
Parameter	M_3			M_4			M_5		
	Mean	5%	95%	Mean	5%	95%	Mean	5%	95%
fico	-0.700	-0.820	-0.581	-0.716	-0.842	-0.589	-0.701	-0.821	-0.581
cltv	0.517	0.337	0.705	0.477	0.290	0.667	0.516	0.333	0.703
orig_upb	-0.156	-0.297	-0.015	-0.088	-0.233	0.060	-0.155	-0.300	-0.014
dti	0.151	0.018	0.282	0.150	0.016	0.282	0.152	0.021	0.283
n_borr	-0.276	-0.536	-0.014	-0.288	-0.548	-0.029	-0.270	-0.527	-0.018
loan_purpose	-0.969	-1.239	-0.695	-0.969	-1.248	-0.686	-0.971	-1.246	-0.696
λ_f	1.165	0.781	1.572	3.555	2.587	4.541	1.317	0.895	1.771
β_{01}	-0.444	-0.465	-0.422	-0.278	-0.292	-0.264	-0.280	-0.294	-0.266
σ_{U_0}	1.860	1.836	1.885	1.236	1.218	1.254	1.237	1.219	1.255
σ	0.734	0.732	0.736	0.706	0.704	0.708	0.706	0.704	0.708
ϕ				0.357	0.354	0.361	0.357	0.353	0.360
σ_{U_1}	0.086	0.085	0.088	0.053	0.052	0.054	0.053	0.052	0.054
ρ_U	-0.782	-0.790	-0.775	-0.810	-0.817	-0.803	-0.811	-0.818	-0.804

Table 3.5: Summary of the posterior distributions of each model's parameters with fold one kept out.

		$\Delta\widehat{AUC}_c^{12}$				
Time(c)	$\widehat{AUC}_c^{12}$ M_0	M_1	M_2	M_3	M_4	M_5
6	0.732	**0.068 (0.046)**	0.067 (0.045)	0.021 (0.028)	-0.010 (0.020)	0.021 (0.029)
7	0.750	0.050 (0.052)	**0.050 (0.050)**	0.014 (0.028)	-0.024 (0.024)	0.013 (0.036)
8	0.796	0.025 (0.017)	**0.025 (0.017)**	-0.003 (0.008)	-0.059 (0.010)	-0.013 (0.005)
9	0.792	**0.010 (0.017)**	0.010 (0.016)	-0.008 (0.020)	-0.034 (0.011)	-0.005 (0.006)
10	0.791	0.004 (0.009)	0.004 (0.009)	-0.012 (0.027)	-0.012 (0.017)	**0.007 (0.008)**
11	0.799	0.001 (0.013)	0.001 (0.013)	-0.008 (0.033)	-0.013 (0.029)	**0.011 (0.018)**
12	0.790	**0.009 (0.015)**	0.008 (0.015)	-0.011 (0.030)	-0.032 (0.033)	-0.003 (0.023)
13	0.794	0.005 (0.018)	0.004 (0.017)	**0.006 (0.023)**	-0.009 (0.023)	-0.011 (0.023)
14	0.778	0.002 (0.015)	0.002 (0.014)	0.041 (0.043)	**0.042 (0.044)**	0.000 (0.026)
15	0.785	-0.002 (0.013)	-0.003 (0.013)	0.046 (0.040)	**0.060 (0.040)**	0.004 (0.030)
16	0.783	-0.006 (0.014)	-0.007 (0.013)	0.050 (0.026)	**0.075 (0.026)**	0.006 (0.022)
17	0.779	-0.009 (0.012)	-0.010 (0.011)	0.057 (0.036)	**0.089 (0.033)**	0.018 (0.033)
18	0.768	-0.008 (0.013)	-0.008 (0.012)	0.066 (0.035)	**0.105 (0.031)**	0.029 (0.034)
19	0.767	-0.006 (0.011)	-0.007 (0.011)	0.061 (0.031)	**0.105 (0.028)**	0.032 (0.031)
20	0.762	-0.009 (0.011)	-0.009 (0.011)	0.060 (0.031)	**0.111 (0.027)**	0.035 (0.034)
21	0.774	-0.006 (0.011)	-0.007 (0.010)	0.054 (0.038)	**0.104 (0.033)**	0.037 (0.039)
22	0.761	-0.006 (0.012)	-0.006 (0.012)	0.069 (0.051)	**0.123 (0.046)**	0.055 (0.052)
23	0.750	-0.007 (0.010)	-0.008 (0.009)	0.073 (0.049)	**0.132 (0.045)**	0.062 (0.052)
24	0.757	-0.016 (0.005)	-0.016 (0.004)	0.062 (0.044)	**0.124 (0.044)**	0.053 (0.047)

Table 3.6: Mean difference of $\widehat{AUC}_c^{\Delta c}$ (Equation 3.12) with respect to model M_0 (Cox model) and prediction window of 12 months ($\Delta c = 12$). The Time(c) column represents c, the known history when making the prediction. The number in parentheses is the standard deviation of the cross-validation analysis (corrected by the overlapping training sets). The largest increment of the corresponding row is marked in bold.

		$\Delta\widehat{PE}_c^{12}$				
Time(c)	$\widehat{PE}_c^{12}$ M_0	M_1	M_2	M_3	M_4	M_5
6	0.367	-0.021 (0.009)	**-0.022 (0.009)**	-0.020 (0.008)	-0.017 (0.008)	-0.018 (0.008)
7	0.397	-0.019 (0.007)	**-0.020 (0.007)**	-0.018 (0.007)	-0.017 (0.007)	-0.017 (0.006)
8	0.428	-0.014 (0.008)	-0.014 (0.009)	-0.015 (0.009)	**-0.017 (0.009)**	-0.015 (0.009)
9	0.467	-0.010 (0.006)	-0.010 (0.006)	-0.009 (0.006)	**-0.014 (0.006)**	-0.011 (0.006)
10	0.487	-0.007 (0.004)	-0.008 (0.004)	0.002 (0.007)	**-0.012 (0.004)**	-0.009 (0.004)
11	0.530	-0.006 (0.003)	-0.006 (0.003)	0.032 (0.022)	**-0.008 (0.003)**	-0.007 (0.003)
12	0.590	-0.005 (0.002)	**-0.005 (0.002)**	0.089 (0.022)	0.010 (0.007)	-0.004 (0.002)
13	0.617	-0.003 (0.001)	**-0.004 (0.001)**	0.168 (0.059)	0.071 (0.040)	0.002 (0.003)
14	0.680	-0.003 (0.001)	**-0.004 (0.001)**	0.373 (0.135)	0.337 (0.131)	0.022 (0.013)
15	0.744	-0.002 (0.002)	**-0.002 (0.002)**	0.516 (0.250)	0.671 (0.322)	0.054 (0.038)
16	0.805	-0.001 (0.003)	**-0.001 (0.002)**	0.601 (0.315)	1.022 (0.471)	0.103 (0.073)
17	0.806	0.000 (0.003)	**-0.001 (0.003)**	0.668 (0.357)	1.389 (0.605)	0.161 (0.108)
18	0.851	0.000 (0.003)	**-0.001 (0.003)**	0.730 (0.368)	1.789 (0.710)	0.230 (0.141)
19	0.911	0.000 (0.003)	0.000 (0.003)	0.691 (0.364)	1.968 (0.807)	0.257 (0.162)
20	0.918	0.000 (0.004)	**-0.001 (0.003)**	0.646 (0.342)	2.068 (0.885)	0.291 (0.173)
21	0.951	0.000 (0.004)	0.000 (0.004)	0.591 (0.258)	2.135 (0.762)	0.305 (0.150)
22	0.916	0.000 (0.004)	0.000 (0.004)	0.468 (0.245)	1.889 (0.810)	0.264 (0.155)
23	0.919	0.002 (0.002)	0.002 (0.002)	0.386 (0.138)	1.783 (0.610)	0.247 (0.103)
24	0.908	0.004 (0.002)	0.003 (0.003)	0.373 (0.127)	1.720 (0.553)	0.269 (0.116)

***For ease of visualisation, all values are multiplied by 100.**

Table 3.7: Mean difference of $\widehat{PE}_c^{\Delta c}$ (Equation 3.14) with respect to model M_0 (Cox model) and prediction window of 12 months ($\Delta c = 12$). The Time(c) column represents c, the known history when making the prediction. The number in parentheses is the standard deviation of the cross-validation analysis (corrected by the overlapping training sets). The largest reduction of the corresponding row is marked in bold.

Chapter 4

Joint Model of Multivariate Longitudinal Outcomes

This chapter is based on a manuscript submitted to a peer-reviewed journal and is currently under "revise and resubmit".

We presented the joint model with one longitudinal outcome in the previous chapter. We were interested in including autoregressive terms and exploring how this extension influences the model's performance. Instead, this chapter focuses on situations where we have more than one longitudinal outcome and how this multivariate version can be efficiently estimated to scale it to credit-related applications.

The chapter is organised in the following way. In Section 4.1, we contextualise the contributions and present the relevant literature. Section 4.2 presents the joint model with multivariate longitudinal outcomes and discrete survival time. We also introduce the estimation procedure using INLA and our proposal to compute the individual survival predictions. Moreover, in Section 4.3, we study how the estimation via INLA methodology behaves under different scenarios through a simulation study. In Section 4.4, we apply two multivariate joint models to predict repayment behaviour in a German consumer loan portfolio and compare them with standard approaches in credit risk analysis. A discussion concludes the chapter.

4.1 Introduction

As discussed earlier in Chapter 2, survival analysis is a convenient approach when we are interested in predicting when the event will occur in the presence of censoring. Survival analysis has been widely applied in the credit risk literature over the years and remains an active area of research (see Banasik et al., 1999; Bellotti and Crook, 2009a; Leow and Crook, 2016; Wang et al., 2020; Blumenstock et al., 2022). However, there has been little effort to address how time-varying covariates (TVCs) that are endogenous to the time of the borrower's event should be included in survival models. The standard practice is to include the observed values of these TVCs so that partial likelihood estimation can be carried out. This practice, however, assumes that the stochastic nature of the TVCs can be disentangled from the survival process. In other words, it treats them as exogenous, which can lead to biased estimators (Kalbfleisch and Prentice, 2002; Rizopoulos, 2012), but also this practice lacks a prediction framework that accounts for the mutual evolution between the TVCs and survival processes.

In addition, another common practice for incorporating TVCs in a survival model and making predictions is to include them with their lagged values, which may reduce the endogenous effect but incorporates other limitations. First, this practice forces us to remove the first lagged observations and therefore assumes that there were no events in that period, hence inducing bias. This period is usually 12 months for credit risk applications. Second, since the event at a particular time is related to observations in previous months, this relationship might not be optimal for predictive purposes. And finally, the lag is usually decided regarding the prediction window, limiting the analysis to other time horizons (Bellotti and Crook, 2013)[1].

We introduced joint models as a flexible and attractive alternative to not only take care of the endogeneity problem that may exist but also because it presents a natural dynamic prediction framework that does not have to make use of the practices mentioned above (Wu and Carroll, 1988; Tsiatis and Davidian, 2004; Henderson et al., 2000; Rizopoulos, 2012). However, we also noted that its flexibility is highly constrained by its estimation cost, which is further increased if

[1]Strictly speaking, we could also predict beyond the prediction window by forecasting the TVCs first, but this is precisely what joint models do without the need to have two separate predictions.

more data and covariates are considered. For this reason and the lack of adequate software (Furgal et al., 2019), most of the joint model literature focuses on the case with only one longitudinal outcome (Hickey et al., 2016) and or relatively small datasets (Rizopoulos and Ghosh, 2011; Brown et al., 2005; Chi and Ibrahim, 2006).

In credit analysis, though, datasets are usually fairly large, in the order of magnitude of ten thousand or one hundred thousand for practical applications. In our empirical case shown in Section 4.4, for instance, we have approximately 60,000 observations. Moreover, common datasets have many time-fixed covariates recorded at the time of origination (e.g. amount of the loan, interest rate, term, among others), and more than one TVC, for example, balance of the loan, number of instalments in arrears, among others. To study the full potential of joint models in credit-related applications, we need fast inference methods that can handle more than one TVC (multivariate) and can scale to large samples. Furthermore, most joint models are implemented assuming time as continuous. Still, loan data are typically delivered over discrete periods (e.g. monthly accounting data), and many ties occur between events making the discrete-time version more appropriate (Bellotti and Crook, 2013; Djeundje and Crook, 2019b).

The application that motivates this chapter corresponds to the prediction of when and which borrower will repay its loan before the date agreed in the contract and in the presence of endogenous TVCs. The prepayment risk concerns banks since it results in unscheduled cash inflows and potential loss of interest. Hence, implementing models that could accurately predict this risk can help banks in their decision-making (BCBS, 2019). The dataset is provided by a bank and involves consumer loans granted in Germany.

We present two methodological and two empirical contributions to the literature. From the methodological point of view, first, we propose a joint model for bivariate longitudinal outcomes and discrete survival data using integrated nested Laplace approximations (INLA) (Rue et al., 2009), a deterministic algorithm for Bayesian inference. We extend Van Niekerk et al. (2019) who use INLA to estimate a joint model for the univariate case with continuous time. By using this method, we suggest a faster estimation procedure that can effortlessly scale to large datasets without compromising the accuracy of the estimates and that otherwise would not be computationally feasible. We illustrate the implementation

via simulation analysis that shows a satisfactory recovery of the true parameter values. Second, we propose a methodology for individual survival predictions using the Laplace method (Tierney and Kadane, 1986) that leads to more accurate approximations than the empirical Bayes approach used in the joint model literature (Rizopoulos, 2012).

From the empirical perspective, our first contribution is applying a multivariate joint model approach in the credit risk context for the first time, particularly for predicting the probability of full prepayment in a consumer loan portfolio. While Hu and Zhou (2019) use joint models to predict early mortgage loan repayment events and show performance improvements compared to survival models, the authors consider only the univariate case, time as continuous and few predictive time horizons in their analysis. Second, we show that these multivariate approaches result in better discrimination and calibration than the survival models commonly used in the literature (Thomas et al., 2017).

4.2 Methodology

4.2.1 Multivariate joint model

For borrower i $(i = 1, \ldots, N)$ we are interested in modelling the time to event T_i, whose domain belongs to positive integer values, in terms of a vector of fixed covariates $\boldsymbol{z}_i$ and a set of M longitudinal outcomes $Y_{i,s}^{(m)}$ $(m = 1, \ldots, M)$ observed at time s $(s = 1, \ldots, T)$.

We assume the length of the study is T, and we observe subject i until time t_i (i.e. $t_i \leq T$ and $s \in \{1, .., t_i\}$), at which point either the event happens, or it is right-censored. In principle, the number of observed measurements for the longitudinal outcomes can differ from the number of survival points. However, we usually have correlative monthly observations in credit-related datasets with no missing values. The reason is that the institution is responsible and interested in keeping track of the credit performance. Therefore, we consider we have t_i measurements for each longitudinal outcome.

The m-th longitudinal outcome $Y_{i,s}^{(m)}$ is assumed as a noisy version of an underlying latent predictor $\eta_{Yi,s}^{(m)}$ that can be decomposed into fixed and random effects. The fixed effects are represented by $\boldsymbol{q}_{i,s}^{(m)\intercal}\boldsymbol{\beta}_1^{(m)}$, where $\boldsymbol{q}_{i,s}^{(m)}$ is a vector of

longitudinal covariates measured at time s with corresponding coefficients $\boldsymbol{\beta}_1^{(m)}$. The random effects are $\boldsymbol{d}_{i,s}^{(m)\intercal}\boldsymbol{U}_i$, where $\boldsymbol{d}_{i,s}^{(m)}$ is the design vector at time s and $\boldsymbol{U}_i$ the corresponding borrower-specific random effects. This leads to the following mixed-effect model (Laird and Ware, 1982) for each longitudinal outcome

$$\begin{aligned}(Y_{i,s}^{(m)}|\eta_{Yi,s}^{(m)},\tau^{(m)}) \sim N(\eta_{Yi,s}^{(m)}, 1/\tau^{(m)}), \qquad m = 1,..,M \\ \eta_{Yi,s}^{(m)} = \boldsymbol{q}_{i,s}^{(m)\intercal}\boldsymbol{\beta}_1^{(m)} + \boldsymbol{d}_{i,s}^{(m)\intercal}\boldsymbol{U}_i,\end{aligned} \tag{4.1}$$

where $\tau^{(m)}$ is the precision of the error terms associated with the m-th longitudinal outcome. The random effects $\boldsymbol{U}_i$ are assumed as mutually independent and distributed as a zero-mean multivariate Gaussian distribution with $r \times r$ precision matrix $Q_{\boldsymbol{U}}$. Note that for the particular case of the mixed-effect "intercept-and-slope" model, we write $\boldsymbol{q}_{i,s}^{(m)\intercal}\boldsymbol{\beta}_1^{(m)} = \beta_{01}^{(m)} + s \cdot \beta_{11}^{(m)}$ and $\boldsymbol{d}_{i,s}^{(m)\intercal}\boldsymbol{U}_i = U_{0i}^{(m)} + s \cdot U_{1i}^{(m)}$ with $\boldsymbol{U}_i = [U_{0i}^{(1)}, U_{1i}^{(1)}, \ldots, U_{0i}^{(M)}, U_{1i}^{(M)}]^\intercal$.

For the survival process, and since we are interested in its discrete version, we keep our representation of the survival points through a binary random variable $X_{i,s}$ that takes the value 1 if borrower i experiences the event at time s and 0 otherwise (Allison, 1982). Hence, the last observation of the sequence for borrower i, i.e. x_{i,t_i} (as before, we denote the realisations of a random variable in lowercase), is equal to the event indicator. Considering the logit link between the binary random variable $X_{i,s}$ and the linear predictor $\eta_{Xi,s}$, the discrete-time survival follows

$$\begin{aligned}(X_{i,s}|X_{i,s-1} = 0, \eta_{Xi,s}) \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{Xi,s})) \\ \eta_{Xi,s} = \nu_s + \boldsymbol{z}_i^\intercal\boldsymbol{\beta}_2 + \sum_{m=1}^{M} \lambda^{(m)} f^{(m)}(\eta_{Yi,s}^{(m)}),\end{aligned} \tag{4.2}$$

where ν_s represents the baseline discrete event time distribution. As we saw previously in Chapter 3, this term is commonly represented by either fixed effects or spline models (Tutz and Schmid, 2016; Djeundje and Crook, 2018). However, in this work, we follow the smoothing approach from Lindgren and Rue (2008) that is implemented in the `INLA` package (Rue et al., 2009)[2]. Specifically, we express ν_s as a discrete time second-order random walk model which is a discretization of a continuous time integrated Wiener process. This smoothing approach depends on one hyperparameter τ_ν that denotes the precision of the underlying Gaussian white noise. Furthermore, $\boldsymbol{\beta}_2$ is the vector of coefficients for the fixed covariates $\boldsymbol{z}_i$. $\lambda^{(m)}$ is the association parameter that links the m-th longitudinal outcome

[2] `INLA` package is hosted on `http://www.r-inla.org/`.

with the survival process. The association function $f^{(m)}(\cdot)$ takes as argument the m-th latent linear predictor $\eta^{(m)}_{Yi,s}$ and returns some of its components or a function of them. For example, a widely used version of the function $f^{(m)}$ is the identity function, i.e. $f^{(m)}(\eta^{(m)}_{Yi,s}) = \eta^{(m)}_{Yi,s}$ (see Hickey et al. (2016) for a comprehensive review of different association functions $f^{(m)}$).

4.2.2 Estimation

Denote the complete observed data as $\mathcal{D} = (\{\boldsymbol{y}_i\}_{i=1,\ldots,N}, \{\boldsymbol{x}_i\}_{i=1,\ldots,N})$, where $\boldsymbol{y}_i$ is the complete history of the M longitudinal outcomes for the borrower i, i.e. $\boldsymbol{y}_i = \{y^{(m)}_{i,s} : s = 1,\ldots,t_i; m = 1,\ldots,M\}$ and $\boldsymbol{x}_i$ is the complete sequence of the binary survival times for borrower i, i.e. $\boldsymbol{x}_i = \{x_{i,s} : s = 1,\ldots,t_i\}$. From Section 4.2.1, the unknown parameters are the coefficients of the longitudinal outcomes $\{\boldsymbol{\beta}^{(m)}_1\}$, the precisions of the error terms $\{\tau^{(m)}\}$, the precision matrix of the random effects $Q_{\boldsymbol{U}}$, the coefficients of the survival process $\boldsymbol{\beta}_2$, the association parameters $\{\lambda^{(m)}\}$, the baseline hazard coefficients $\{\nu_s\}$ and the precision parameter τ_ν.

As seen in Section 2.2, one crucial assumption in the joint model approach is the conditional independence between the survival and longitudinal processes given the random effects $\boldsymbol{U}_i$ (Wu and Carroll, 1988; Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Tsiatis and Davidian, 2004). Thus, the observation density can be easily formulated given the random effects and following, for instance, simulation-based MCMC schemes to estimate the parameters. That is the strategy we follow in Chapter 3. For the multivariate case, the generalisation is straightforward (see Andrinopoulou et al., 2014).

However, this estimation strategy is already computationally expensive for the univariate case or even infeasible for some applications with large sample sizes. A faster and accurate alternative is to use the INLA methodology proposed by Rue et al. (2009) and implemented in the `INLA` package for the `R` software. INLA approximates the Bayesian inference on the class of Latent Gaussian models (LGMs), as presented in Rue et al. (2009) (see Section 2.2.5). This class comprises numerous well-known statistical models, for example, mixed-effects, dynamic, and spatial-temporal models. Our multivariate joint model can also be formulated as an LGM, as shown in Section 2.2.4 for the univariate case.

Using the INLA notation presented in Section 2.2.4, we define the following terms

$$\begin{aligned}\boldsymbol{\mu} &= (\{\boldsymbol{\eta}_{Yi}^{(m)}\}, \{\boldsymbol{\eta}_{Xi}\}, \{\boldsymbol{\beta}_1^{(m)}\}, \{\boldsymbol{U}_i\}, \{\nu_s\}, \boldsymbol{\beta}_2)\\ \boldsymbol{\theta}_1 &= (\theta_{\boldsymbol{\beta}_1^{(m)}}, \tau_\nu, \theta_{\boldsymbol{\beta}_2}, Q_{\boldsymbol{U}}, \{\lambda^{(m)}\})\\ \boldsymbol{\theta}_2 &= (\{\tau^{(m)}\})\end{aligned}$$

where $\theta_{\boldsymbol{\beta}_1^{(m)}}$ and $\theta_{\boldsymbol{\beta}_2}$ are hyperparameters for $\{\boldsymbol{\beta}_1^{(m)}\}$ and $\boldsymbol{\beta}_2$, respectively. Note that with this notation, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ correspond to the set of hyperparameters of the latent field $\boldsymbol{\mu}$ and likelihood, respectively.

Given the conditional dependency assumed in Equations 4.1 and 4.2, the joint conditional density of $\mathcal{D}$ is $p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\theta}_2) = \prod_{j\in\mathcal{J}} p(\mathcal{D}_j|\mu_j, \boldsymbol{\theta}_2)$, where $p(\cdot)$ denotes either a probability mass function or a probability density, as appropriate for each variable, and $\mathcal{J}$ corresponds to the set of indices for all observed values in $\mathcal{D}$, and it is coded so that each observation is associated with its respective linear predictor η (see Section 2.2.5).

According to INLA methodology, the density of $\boldsymbol{\mu}|\boldsymbol{\theta}_1$ is assumed as zero-mean Gaussian with precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_1)$. Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, then the joint posterior distribution follows

$$\begin{aligned}p(\boldsymbol{\mu}, \boldsymbol{\theta}|\mathcal{D}) &\propto p(\boldsymbol{\theta}) p(\boldsymbol{\mu}|\boldsymbol{\theta}) \prod_{j\in\mathcal{J}} p(\mathcal{D}_j|\mu_j, \boldsymbol{\theta})\\ &\propto p(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\boldsymbol{\mu}^\intercal \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{\mu} + \sum_{j\in\mathcal{J}} \log\{p(\mathcal{D}_j|\mu_j, \boldsymbol{\theta})\}\right].\end{aligned}$$

However, we are not interested in explicitly estimating the joint posterior distribution, but rather the posterior marginals, $p(\mu_j|\mathcal{D})$ and $p(\theta_j|\mathcal{D})$, specified by

$$\begin{aligned}p(\mu_j|\mathcal{D}) &= \int p(\mu_j|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}\\ p(\theta_j|\mathcal{D}) &= \int p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}_{-j}.\end{aligned} \tag{4.3}$$

The INLA methodology computes these marginals based on the Laplace approximation (Tierney and Kadane, 1986) over the terms $p(\boldsymbol{\theta}|\mathcal{D})$ and $p(\mu_j|\boldsymbol{\theta}, \mathcal{D})$, as detailed in Section 2.2.5.

To fully specify the estimation procedure, we need to define the priors of the hyperparameters $\boldsymbol{\theta}$. In particular, we assume that the parameters $\{\boldsymbol{\beta}_1^{(m)}\}$ and $\boldsymbol{\beta}_2$ have independent zero-mean Gaussian priors with precision matrix of $\theta_{\boldsymbol{\beta}}\boldsymbol{I}$,

where $\boldsymbol{I}$ is the identity matrix with the corresponding dimension for each set of parameters and $\theta_{\boldsymbol{\beta}}$, a precision parameter, is equal to 0.01 (i.e. $\theta_{\boldsymbol{\beta}_1^{(m)}}$ and $\theta_{\boldsymbol{\beta}_2}$ are all equal to $\theta_{\boldsymbol{\beta}}$). Moreover, for the log scale of $\{\tau^{(m)}\}$, the precision parameters of the error terms of the longitudinal outcomes, we assume weakly informative log-gamma prior distributions with shape and scale parameters of 1 and 5×10^{-5}, respectively. The prior of the $p \times p$ precision matrix $Q_{\boldsymbol{U}}$ is assumed as a Wishart distribution $\mathcal{W}_p(\boldsymbol{I}, p(p+1)/2+1)$ which shows sensible results on the simulation study (see Section 4.3).

Finally, for the prior of the hyperparameter τ_ν that represents the precision of the second-order random walk model, we assume a penalising complexity (PC) prior as described in Simpson et al. (2017). The shape of this prior is defined via the influence of the parameter on the latent process model, as measured by the deviation from a base model with zero variance. The prior is specified by choosing an upper α-quantile u for the standard deviation of the model, so that $P(\tau_\nu^{-1/2} = \sigma_\nu > u) = \alpha$ for the choice of α and u. We use a weakly informative prior by choosing $P(\sigma_\nu > 1) = 0.01$, indicating a small probability for a large standard deviation.

4.2.3 Individual survival prediction

We are interested in estimating how likely the full prepayment event is for a new borrower k not included in the training data $\mathcal{D}$. Analogous to Chapter 3, we assume that we have collected the M longitudinal outcomes for this borrower up to time c. Since we know with certainty that this borrower has not experienced the event until at least c, our interest is in estimating the probability of surviving to time $c + \Delta c$, with $\Delta c \in \mathbb{Z}_+$, conditional on having survived to c. Denote the set of observed measurements for the M longitudinal outcomes as $\boldsymbol{y}_k = \{y_{k,s}^{(m)} : s = 1, \ldots, c; m = 1, \ldots, M\}$, then the conditional probability we are interested in is the following

$$P(T_k > c + \Delta c | T_k > c, \boldsymbol{y}_k, \mathcal{D}) = \int P(T_k > c + \Delta c | T_k > c, \boldsymbol{y}_k, \Theta) p(\Theta | \mathcal{D}) \mathrm{d}\Theta, \quad (4.4)$$

where Θ represents the set of parameters to estimate for the model specification i.e. $\Theta = \{\{\boldsymbol{\beta}_1^{(m)}\}, \{\nu_s\}, \boldsymbol{\beta}_2, \{\lambda^{(m)}\}, \tau_\nu, Q_{\boldsymbol{U}}, \{\tau^{(m)}\}\}$, and $p(\Theta|\mathcal{D})$ the corresponding posterior distribution. Note that we explicitly left the random effects of borrower k, $\boldsymbol{U}_k$, out of Θ since k is not included in $\mathcal{D}$, so we do not have estimation of them.

However, as we will see below, to evaluate Equation 4.4 and take advantage of the conditional independence assumption, we need to marginalised over $\boldsymbol{U}_k$. In any case, Equation 4.4 does not have a closed form and needs to be numerically estimated.

Monte Carlo simulation schemes have been proposed to estimate Equation 4.4 (Rizopoulos, 2011; Proust-Lima and Taylor, 2009). Yet, in credit-related applications, the interest is in estimating predictions for many out-of-sample borrowers, and these simulation schemes are computationally expensive. Therefore, we propose the following procedure to approximate Equation 4.4.

The first step considers that we have already estimated the posterior distribution of Θ, and we can rely on a point estimate denoted by $\hat{\Theta}$ (we use the posterior mean). Thus, expression $P(T_k > c + \Delta c | T_k > c, \boldsymbol{y}_k, \hat{\Theta})$ can be further marginalised over the random effects $\boldsymbol{U}_k$ as follows

$$P(T_k > c + \Delta c | T_k > c, \boldsymbol{y}_k, \hat{\Theta}) = \int P(T_k > c + \Delta c | T_k > c, \boldsymbol{U}_k, \hat{\Theta}) \times p(\boldsymbol{U}_k | T_k > c, \boldsymbol{y}_k, \hat{\Theta}) \mathrm{d}\boldsymbol{U}_k, \tag{4.5}$$

where $P(T_k > c + \Delta c | T_k > c, \boldsymbol{U}_k, \hat{\Theta})$ can be in turn written as

$$\begin{aligned} P(T_k > c + \Delta c | T_k > c, \boldsymbol{U}_k, \hat{\Theta}) &= \frac{P(T_k > c + \Delta c | \boldsymbol{U}_k, \hat{\Theta})}{P(T_k > c | \boldsymbol{U}_k, \hat{\Theta})} \\ &= \frac{\prod_{s=1}^{c+\Delta c}(1 - p_{k,s})}{\prod_{s=1}^{c}(1 - p_{k,s})} = \prod_{s=c+1}^{c+\Delta c}(1 - p_{k,s}), \end{aligned}$$

with $p_{k,s} = \text{logit}^{-1}(\eta_{Xk,s})$ and $\eta_{Xk,s}$ following Equation 4.2.

A first order approximation of Equation 4.5 is presented in Rizopoulos (2012) who uses the empirical Bayes estimates for the random effects $\boldsymbol{U}_k$ as follows

$$P(T_k > c + \Delta c | T_k > c, \boldsymbol{y}_k, \hat{\Theta}) = \frac{P(T_k > c + \Delta c | \hat{\boldsymbol{U}}_k, \hat{\Theta})}{P(T_k > c | \hat{\boldsymbol{U}}_k, \hat{\Theta})} + O\left(\frac{1}{c}\right)$$

where $\hat{\boldsymbol{U}}_k = \text{argmax}_{\boldsymbol{U}}\{\log P(T_k > c, \boldsymbol{y}_k, \boldsymbol{U} | \hat{\Theta})\}$. This is the approach we follow in Chapter 3.

However, we can get a better approximation of the Equation 4.5 by using the Laplace method introduced by Tierney and Kadane (1986). First note that $P(\boldsymbol{U} | T_k > c, \boldsymbol{y}_k, \hat{\Theta})$ can be expressed as $P(T_k > c, \boldsymbol{y}_k, \boldsymbol{U} | \hat{\Theta}) / P(T_k > c, \boldsymbol{y}_k | \hat{\Theta})$ and

the term $P(T_k > c, \boldsymbol{y}_k|\hat{\Theta})$ can be marginalised as $\int P(T_k > c, \boldsymbol{y}_k, \boldsymbol{U}|\hat{\Theta})\mathrm{d}\boldsymbol{U}$. This lets us write the following expression

$$P(T_k > c+\Delta c|T_k > c, \boldsymbol{y}_k, \hat{\Theta}) = \frac{\int P(T_k > c+\Delta c|T_k > c, \boldsymbol{U}, \hat{\Theta})P(T_k > c, \boldsymbol{y}_k, \boldsymbol{U}|\hat{\Theta})\mathrm{d}\boldsymbol{U}}{\int P(T_k > c, \boldsymbol{y}_k, \boldsymbol{U}|\hat{\Theta})\mathrm{d}\boldsymbol{U}}.$$

Define $-c \cdot h_k(\boldsymbol{U}) = \log\{P(T_k > c, \boldsymbol{y}_k, \boldsymbol{U}|\hat{\Theta})\}$ and $g(\boldsymbol{U}) = P(T_k > c+\Delta c|T_k > c, \boldsymbol{U}, \hat{\Theta})$. This brings us to the following

$$P(T_k > c+\Delta c|T_k > c, \boldsymbol{y}_k, \hat{\Theta}) = \frac{\int g(\boldsymbol{U})\exp\{-c \cdot h_k(\boldsymbol{U})\}\mathrm{d}\boldsymbol{U}}{\int \exp\{-c \cdot h_k(\boldsymbol{U})\}\mathrm{d}\boldsymbol{U}}. \tag{4.6}$$

Since $g(\boldsymbol{U}) > 0$, using the Laplace method (Tierney and Kadane, 1986), we can approximate the last expression as

$$P(T_k > c+\Delta c|T_k > c, \boldsymbol{y}_k, \hat{\Theta}) = \frac{|\Sigma^*|^{1/2}\exp\{-c \cdot h_k^*(\boldsymbol{U}_k^*)\}}{|\tilde{\Sigma}|^{1/2}\exp\{-c \cdot h_k(\hat{\boldsymbol{U}}_k)\}} + O\left(\frac{1}{c^2}\right) \tag{4.7}$$

where $-c \cdot h_k^*(\boldsymbol{U}) = -c \cdot h_k(\boldsymbol{U}) + \log g(\boldsymbol{U})$. The vectors $\boldsymbol{U}_k^*$ and $\hat{\boldsymbol{U}}_k$ are the arguments of the maxima of $-h_k^*(\cdot)$ and $-h_k(\cdot)$, respectively. Σ^* and $\tilde{\Sigma}$ are the inverse of the Hessians for h_k^* and h_k, respectively, evaluated at $\boldsymbol{U}_k^*$ and $\hat{\boldsymbol{U}}_k$.

Note also that we can recover the approach of Rizopoulos (2012) described as the first-order approximation of Equation 4.6 by applying the Laplace method separately in the numerator and denominator of Equation 4.6. Explicitly,

$$\begin{aligned}
\frac{\int g(\boldsymbol{U})\exp\{-c \cdot h_k(\boldsymbol{U})\}\mathrm{d}\boldsymbol{U}}{\int \exp\{-c \cdot h_k(\boldsymbol{U})\}\mathrm{d}\boldsymbol{U}} &= \frac{g(\hat{\boldsymbol{U}}_k)(2\pi/c)^{p/2}|\tilde{\Sigma}|^{1/2}\exp\{-c \cdot h_k(\hat{\boldsymbol{U}}_k)\}[1+O_1(1/c)]}{(2\pi/c)^{p/2}|\tilde{\Sigma}|^{1/2}\exp\{-c \cdot h_k(\hat{\boldsymbol{U}}_k)\}[1+O_2(1/c)]} \\
&= g(\hat{\boldsymbol{U}}_k)\left[1 + \frac{O_1(1/c) - O_2(1/c)}{1+O_2(1/c)}\right] \\
&= g(\hat{\boldsymbol{U}}_k) + O\left(\frac{1}{c}\right) \\
&= \frac{P(T_k > c+\Delta c|\hat{\boldsymbol{U}}_k, \hat{\Theta})}{P(T_k > c|\hat{\boldsymbol{U}}_k, \hat{\Theta})} + O\left(\frac{1}{c}\right)
\end{aligned}$$

Equation 4.7 provides a second-order approximation to the conditional probability, hence is the procedure we follow in the calculations required by the performance metrics described below.

4.2.4 Performance metrics

We are interested in measuring the performance of the models by their ability to differentiate borrowers who will prepay from those who will not (discrimination)

and by their ability to estimate accurate probabilities (calibration). The metrics we use for discrimination and calibration are similar to those described in Section 3.2.3. Below, we highlight only the most relevant aspects.

For discrimination, we use the concordance index (Harrell et al., 1982) in the version introduced by Rizopoulos (2011) which we adapt here to the discrete form as

$$C_{AUC}^{\Delta c} = \sum_{c} AUC_{c}^{\Delta c} u(c), \tag{4.8}$$

where $u(c)$ is a weight function to account for the fact that not all time points contribute the same. The choice of that function remains an open question. Rizopoulos (2011) proposes the use of $u(c) = P(T_i > c)/\sum_t P(T_i > t)$, where $P(T_i > c)$ is the marginal survival probability that can be estimated with the Kaplan-Meier estimator (Kalbfleisch and Prentice, 2002) and this is the approach that we also follow.

Moreover, as described in Chapter 3, $AUC_c^{\Delta c}$ is the discrete-time dependent AUC with censoring and is decomposed by the sum of four components $\widehat{AUC}_l(c, \Delta c)$ ($l = 1, 2, 3, 4$), each of them representing a possible pair combination between censored and not censored borrowers (set of concordant pairs, represented by $\Omega_{ij}^{(l)}$) in the following way

$$\widehat{AUC}_l(c, \Delta c) = \frac{\sum_i^N \sum_{j \neq i}^N I(\hat{\pi}_i(c + \Delta c|c) < \hat{\pi}_j(c + \Delta c|c)) \cdot I(\Omega_{ij}^{(l)}) \cdot \hat{\nu}_{ij}^{(l)}}{\sum_i^N \sum_{j \neq i}^N I(\Omega_{ij}^{(l)}) \cdot \hat{\nu}_{ij}^{(l)}}.$$

To shorten the notation, we have used $\hat{\pi}_i(c + \Delta c|c)$ to denote the estimated conditional survival probability $P(T_i > c + \Delta c|T_i > c, \boldsymbol{y}_i, \mathcal{D})$ that follows Equation 4.4. $I(\cdot)$ denotes the indicator function and the terms $\hat{\nu}_{ij}^{(l)}$ represent the probability that the pairs are comparable.

For measuring calibration, we follow analogously to Equation 4.8 to assess the overall calibration performance as

$$C_{EPE}^{\Delta c} = \sum_{c} \widehat{EPE}(c + \Delta c|c) u(c), \tag{4.9}$$

where $\widehat{EPE}$ is the estimation of the expected prediction error specified as (Henderson et al., 2002)

$$\widehat{EPE}(c + \Delta c|c) = n(c)^{-1} \sum_{i:T_i > c} \left\{ S_i(c + \Delta c|c) + E_i(c + \Delta c|c) + C_i(c + \Delta c|c) \right\},$$

where $n(c)$ is the number of borrowers at risk at time c, and the other three terms inside the sum are defined as

$$
\begin{aligned}
S_i(c+\Delta c|c) &= I(T_i > c+\Delta c)L\{1, \hat{\pi}_i(c+\Delta c|c)\} \\
E_i(c+\Delta c|c) &= \delta_i I(T_i \leq c+\Delta c)L\{0, \hat{\pi}_i(c+\Delta c|c)\} \\
C_i(c+\Delta c|c) &= (1-\delta_i)I(T_i \leq c+\Delta c)\Big[\hat{\pi}_i(c+\Delta c|T_i)L\{1, \hat{\pi}_i(c+\Delta c|c)\} + \\
&\qquad + (1-\hat{\pi}_i(c+\Delta c|T_i))L\{0, \hat{\pi}_i(c+\Delta c|c)\}\Big],
\end{aligned}
$$

where δ_i is the event indicator and $L(\cdot,\cdot)$ represents the loss function. In this chapter, we use the logarithmic score (Good, 1952) as the loss function instead of the widely used Brier score because it is consistent with the use of likelihoods (or log-likelihoods) to measure the models (Winkler, 1969).

4.3 Simulation

In this section, we perform a simulation study of the discrete multivariate joint model with INLA presented in Section 4.2. The aim is to check how well the proposed implementation works under different sample sizes. The simulated setting is motivated by the application described in Section 4.4. It follows a joint model with two longitudinal outcomes, both of which have a fixed intercept plus random intercept and slope. The four random effects, two intercepts and two slopes, are assumed zero-mean multivariate Gaussian distributed. Moreover, the event process has an additional covariate, fixed in time, and the baseline hazard rate is drawn from a cubic polynomial function. Formally, the generated data for both longitudinal processes follows

$$
\begin{aligned}
(Y_{i,s}^{(m)}|\eta_{Yi,s}^{(m)}, \tau^{(m)}) &\sim N(\eta_{Yi,s}^{(m)}, 1/\tau^{(m)}), \qquad m = 1, 2, \\
\eta_{Yi,s}^{(m)} &= \beta_{01}^{(m)} + U_{0i}^{(m)} + U_{1i}^{(m)} \cdot s, \\
(U_{0i}^{(1)}, U_{1i}^{(1)}, U_{0i}^{(2)}, U_{1i}^{(2)})^{\intercal} &\sim N_4(\mathbf{0}, Q_{\boldsymbol{U}}^{-1}),
\end{aligned}
$$

where covariance matrix $Q_{\boldsymbol{U}}^{-1}$ is parameterised via marginal precisions $\tau_{U_{01}}$, $\tau_{U_{11}}$, $\tau_{U_{02}}$, and $\tau_{U_{12}}$, and pairwise correlations ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24}, and ρ_{34}. Formally,

$$
Q_{\boldsymbol{U}}^{-1} = \begin{pmatrix}
1/\tau_{U_{01}} & \rho_{12}/\sqrt{\tau_{U_{01}}\tau_{U_{11}}} & \rho_{13}/\sqrt{\tau_{U_{01}}\tau_{U_{02}}} & \rho_{14}/\sqrt{\tau_{U_{01}}\tau_{U_{12}}} \\
\rho_{12}/\sqrt{\tau_{U_{01}}\tau_{U_{11}}} & 1/\tau_{U_{11}} & \rho_{23}/\sqrt{\tau_{U_{11}}\tau_{U_{02}}} & \rho_{24}/\sqrt{\tau_{U_{11}}\tau_{U_{12}}} \\
\rho_{13}/\sqrt{\tau_{U_{01}}\tau_{U_{02}}} & \rho_{23}/\sqrt{\tau_{U_{11}}\tau_{U_{02}}} & 1/\tau_{U_{02}} & \rho_{34}/\sqrt{\tau_{U_{02}}\tau_{U_{12}}} \\
\rho_{14}/\sqrt{\tau_{U_{01}}\tau_{U_{12}}} & \rho_{24}/\sqrt{\tau_{U_{11}}\tau_{U_{12}}} & \rho_{34}/\sqrt{\tau_{U_{02}}\tau_{U_{12}}} & 1/\tau_{U_{12}}
\end{pmatrix}.
$$

Moreover, the corresponding event process follows

$$\begin{aligned}(X_{i,s}|X_{i,s-1}=0,\eta_{Xi,s}) &\sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{Xi,s})),\\ \eta_{Xi,s} &= \nu_s+\beta_{12}z_i+\lambda^{(1)}\eta^{(1)}_{Yi,s}+\lambda^{(2)}\eta^{(2)}_{Yi,s},\\ \nu_s &= c_0+c_1s^1+c_2s^2+c_3s^3.\end{aligned} \tag{4.10}$$

We simulate these data considering a maximum of 36 periods for three different numbers of borrowers (500, 1,000 and 1,500) which correspond to 15,183, 30,187 and 44,971 observations, respectively. For illustrative purposes, in Figure 4.1 we show how the simulated events are distributed over time for the sample with 1,500 borrowers.

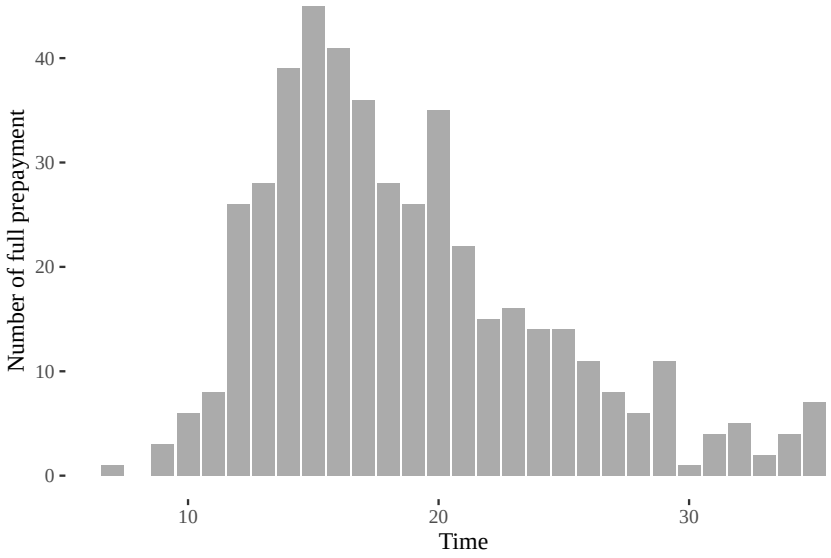

Figure 4.1: Time-events distribution for the simulated sample of 1,500 borrowers.

Similarly, Figure 4.2 shows the evolution in time for both simulated longitudinal outcomes. To aid visualisation, we highlight, in dashed red lines, ten borrowers that experienced the event and, in dotted blue lines, ten that are censored.

As noted in Equation 4.10, the baseline hazard rate ν_s is generated from a cubic polynomial function. However, when estimating the model, we assume that ν_s has a more flexible prior specification than a polynomial function, expressly, a second-order random walk model (Lindgren and Rue, 2008). In Figure 4.3, we show the true baseline hazard rate in solid black line (after the logistic transformation) and the estimated 95% credible intervals for the three samples. We observe a good fit for all three samples, and as we increase the sample size, the interval narrows around the true value.

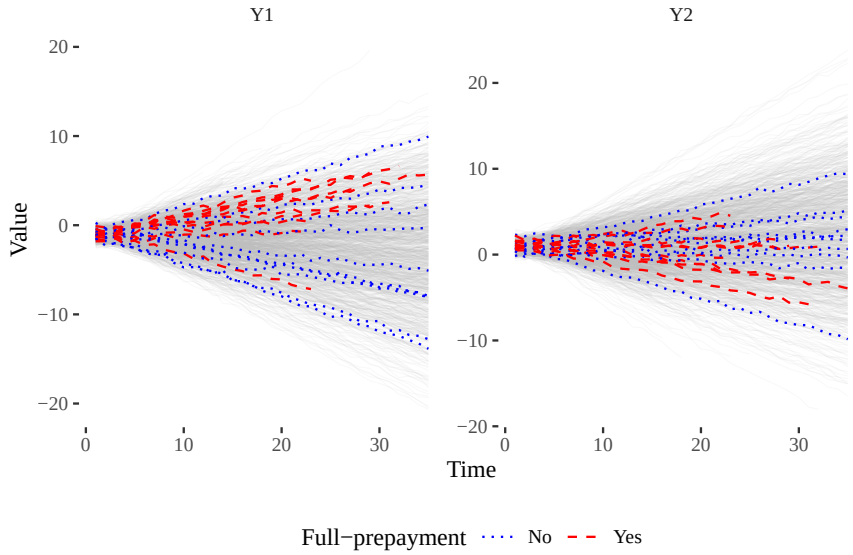

Figure 4.2: Both longitudinal outcomes for the simulated sample of 1,500 borrowers. For visual purposes, we highlight ten borrowers who experienced the event (dashed line) and ten who are censored (dotted line).

Table 4.1 shows the true values of the parameters in the simulation setting and the ones estimated under each sample size. We see a good recovery of the true parameter values for the three sample sizes and, similarly to the behaviour observed in Figure 4.3, as the sample size increases, the credible intervals are narrower.

Finally, to show how precise and fast is the INLA methodology for our model in comparison with an MCMC scheme, in Appendix B.1 we carry out a comparative analysis. We show, among other settings, that the inference for $N = 500$ takes more than 4 hours when performed by MCMC sampling. In contrast, INLA takes less than 3 minutes using the same computational resources.

4.4 Repayment behaviour in German consumer loans

This section presents the application that motivated our proposal of the joint model of multivariate longitudinal outcomes and discrete survival data. In Section 4.4.1, we describe the data provided by a bank and the sampling strategy to validate the performance out-of-sample and out-of-time. Moreover, in Section 4.4.2, we estimate and compare four models to study the advantages of different specifications.

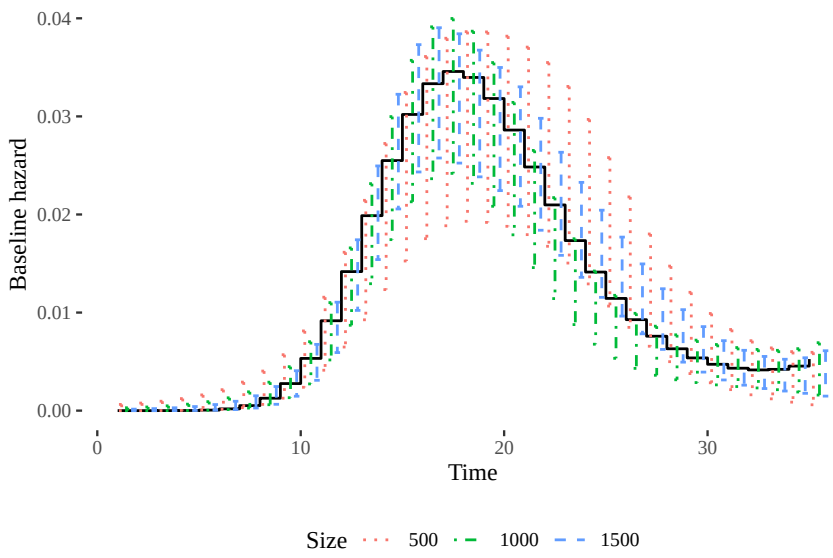

Figure 4.3: Simulated baseline hazard (solid stepped line) and the estimated 95% credible intervals for the three sample sizes.

4.4.1 Data

The complete dataset is formed by two different cohorts of consumer loans granted by a bank. In the first cohort the loans were originated in April, 2012 and in the second the loans were originated in August, 2015. Each cohort has 40 consecutive months of performance. We use the first cohort as the training dataset and the second as the out-of-time dataset.

The training dataset corresponds to 2,397 consumer loans with a total of 59,415 observations. The number of full prepayment events is 470 and its distribution over time is shown in Figure 4.4. The first longitudinal outcome is the cumulative sum of the ratio between the actual balance and the scheduled balance of each loan. This longitudinal outcome accounts for how different the loan balance is from originally scheduled. That gives early signals, for example, if the loan is underpaid or overpaid. The second longitudinal outcome is an internal score calculated by the bank (on a logarithmic scale), which measures the borrower's creditworthiness. Figure 4.5 shows the evolution of the two longitudinal outcomes and, analogously to Figure 4.2, we have highlighted some borrowers that experienced the full prepayment event (dashed line) and borrowers that do not (dotted line). The rationale for choosing these two longitudinal outcomes is that we would expect that a borrower who pays more than what is scheduled (slope below from the diagonal in Figure 4.4) and whose internal score is high would

		$N = 500$			$N = 1000$			$N = 1500$		
	True	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
$\beta_{01}^{(1)}$	-1.00	-1.02	-1.06	-0.98	-0.99	-1.02	-0.97	-1.00	-1.02	-0.98
$\beta_{01}^{(2)}$	1.00	0.98	0.94	1.02	0.99	0.96	1.02	0.99	0.97	1.01
β_{12}	-0.50	-0.60	-0.71	-0.49	-0.49	-0.56	-0.41	-0.53	-0.58	-0.47
$\tau^{(1)}$	25.00	24.97	24.27	25.55	25.13	24.55	25.56	25.13	24.75	25.44
$\tau^{(2)}$	25.00	25.08	24.47	25.65	25.14	24.74	25.61	25.09	24.77	25.51
$\tau_{U_{01}}$	4.00	4.18	3.68	4.73	4.09	3.75	4.46	4.22	3.93	4.52
$\tau_{U_{11}}$	25.00	25.58	22.63	28.91	22.55	20.61	24.49	23.62	22.06	25.24
$\tau_{U_{02}}$	4.00	3.81	3.35	4.30	4.01	3.65	4.35	3.79	3.54	4.05
$\tau_{U_{12}}$	25.00	26.09	23.06	29.40	25.26	23.07	27.41	23.70	22.15	25.34
ρ_{12}	-0.30	-0.20	-0.28	-0.11	-0.29	-0.35	-0.24	-0.25	-0.30	-0.21
ρ_{13}	0.30	0.36	0.28	0.44	0.32	0.26	0.37	0.30	0.26	0.34
ρ_{14}	0.30	0.24	0.15	0.32	0.30	0.24	0.35	0.28	0.23	0.32
ρ_{23}	0.30	0.32	0.24	0.40	0.33	0.27	0.38	0.34	0.29	0.38
ρ_{24}	0.30	0.32	0.23	0.39	0.29	0.24	0.34	0.30	0.26	0.34
ρ_{34}	-0.30	-0.29	-0.37	-0.21	-0.26	-0.32	-0.20	-0.33	-0.38	-0.29
$\lambda^{(1)}$	0.50	0.57	0.49	0.65	0.55	0.49	0.60	0.52	0.48	0.56
$\lambda^{(2)}$	-0.50	-0.55	-0.64	-0.47	-0.50	-0.56	-0.45	-0.48	-0.52	-0.44

Table 4.1: Estimations for the three simulation settings.

have a higher probability of paying the loan in full. This is further confirmed in the empirical results.

To validate and compare the models described in Section 4.4.2, we perform a ten-fold cross-validation analysis. For each validation set (out-of-sample), we assess the performance in terms of the discrimination and calibration metrics described in Section 4.2.4. Moreover, to assess the robustness of the results, we use the out-of-time dataset mentioned above. This dataset corresponds to 2,516 borrowers with a total of 65,928 observations and there are no overlapping times with the data used in the cross-validation.

4.4.2 Models and results

In analysing a full prepayment dataset, a bank is particularly interested in understanding how precise the model is in predicting full prepayment. We estimate four models to investigate the predictive power of the multivariate joint model framework. The first is a discrete survival model where both longitudinal outcomes

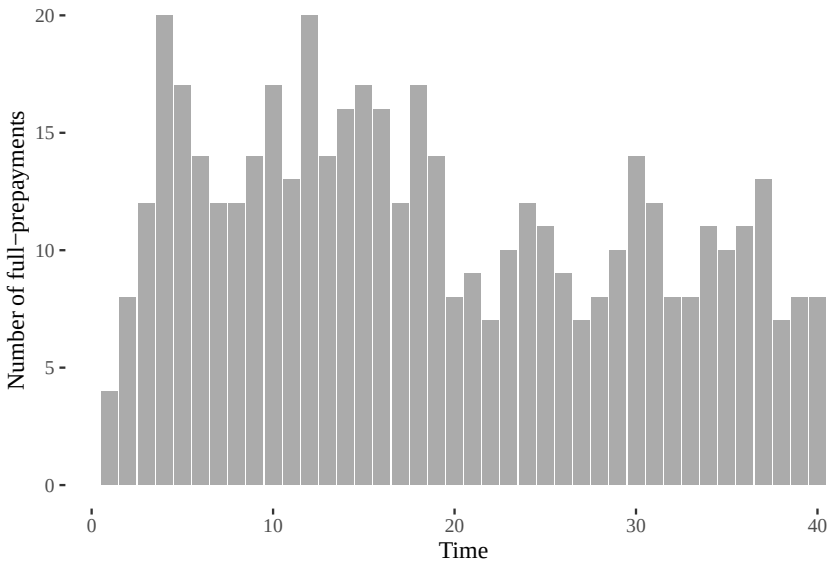

Figure 4.4: Distribution of the full prepayment events in time for the training dataset.

are included as standard TVCs (observed value), so no joint model framework is used. We denote this model as *Cox* model[3]. This model relates the event at month s with the last observation of the longitudinal outcomes. The limitation is that when we are interested in predicting the probability of the event, for example, at $s+12$, we assume the longitudinal outcomes remained constant and equal to the last observed value throughout the prediction window.

The second model *Cox_Lag* is also a discrete survival model. The difference with the *Cox* model is that the event at month t is now related to the observations of the longitudinal outcomes lagged in 12 months. The lag responds to the time window of interest in the predictions, so when we predict the probability of the event at $s+12$, the model is already estimated to consider the observed values at s. These two models, *Cox* and *Cox_Lag*, are the standard survival approaches in credit literature when TVCs are present and, thus, are considered as our natural benchmarks (see, for example, Gross and Souleles, 2002; Bellotti and Crook, 2013; Wang et al., 2020; Calabrese and Crook, 2020).

The third and fourth models (*JM1* and *JM2*, respectively) are both multivariate joint models for discrete survival data. The only difference between them is in the assumed correlations for the random effects. The *JM1* model assumes a correlation between the random effects belonging to each of the longitudinal outcomes but no correlation between the random effects of different longitudinal outcomes.

[3]The model follows the discrete survival approach proposed in Cox (1972).

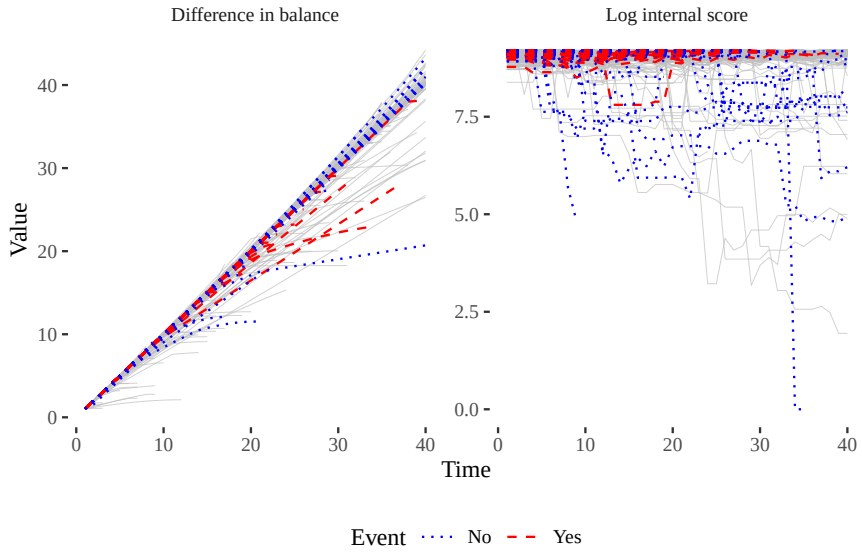

Figure 4.5: Evolution of both longitudinal outcomes for the full prepayment dataset. For visual purposes, we highlight borrowers that full prepaid the loan (dashed line) and borrowers that are censored (dotted line).

The *JM2* model assumes correlation within and between the random effects of both longitudinal outcomes (fully correlated). This last setting investigates if substantial improvements are gained when a more complex relationship between the longitudinal outcomes is used.

Following the notation introduced in Section 4.2, we define $Y_{i,s}^{(1)}$ and $Y_{i,s}^{(2)}$ as the cumulative ratio between the balances and the logarithm of the internal score, respectively, at time s for borrower i. Moreover, we denote $X_{i,s}$ as the binary variable that equals 1 if the borrower i fully prepays the loan at time s and 0 otherwise. $\boldsymbol{z}_i$ is the vector of time-fixed covariates for borrower i (for more details about these covariates, see Appendix B.2) and ν_s is the baseline hazard. The four models' specifications of the event process follow $(X_{i,s}|X_{i,s-1} = 0, \eta_{Xi,s}) \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{Xi,s}))$, the differences come in the assumed predictor $\eta_{Xi,s}$ (Equation 4.2). Moreover, both longitudinal processes assume $(Y_{i,s}^{(m)}|\eta_{Yi,s}^{(m)}, \tau^{(m)}) \sim N(\eta_{Yi,s}^{(m)}, 1/\tau^{(m)})$ for $m = 1, 2$, therefore, we only need to describe the event predictors to fully specify the models. These are the following

Cox. Discrete survival model with TVCs. The event predictor is described as $\eta_{Xi,s} = \nu_s + \boldsymbol{z}_i^\intercal \boldsymbol{\beta}_2 + \lambda^{(1)} y_{i,s}^{(1)} + \lambda^{(2)} y_{i,s}^{(2)}$.

Cox_Lag. Discrete survival model with lagged TVCs. The event predictor is described as $\eta_{Xi,s} = \nu_s + \boldsymbol{z}_i^\intercal \boldsymbol{\beta}_2 + \lambda^{(1)} y_{i,s-12}^{(1)} + \lambda^{(2)} y_{i,s-12}^{(2)}$.

JM1. Joint model not fully correlated. The event predictor is described as $\eta_{Xi,s} = \nu_s + \boldsymbol{z}_i^\intercal \boldsymbol{\beta}_2 + \lambda^{(1)}(U_{0i}^{(1)} + U_{1i}^{(1)} \cdot s) + \lambda^{(2)}(U_{0i}^{(2)} + U_{1i}^{(2)} \cdot s)$ and the corresponding longitudinal processes as

$$\begin{aligned}
\eta_{Yi,s}^{(1)} &= \beta_{01}^{(1)} + \beta_{11}^{(1)} \cdot s + U_{0i}^{(1)} + U_{1i}^{(1)} \cdot s \\
\eta_{Yi,s}^{(2)} &= \beta_{01}^{(2)} + U_{0i}^{(2)} + U_{1i}^{(2)} \cdot s \\
(U_{0i}^{(1)}, U_{1i}^{(1)})^\intercal &\sim N_2(\mathbf{0}, Q_{\boldsymbol{U}_1}^{-1}) \\
(U_{0i}^{(2)}, U_{1i}^{(2)})^\intercal &\sim N_2(\mathbf{0}, Q_{\boldsymbol{U}_2}^{-1}).
\end{aligned}$$

JM2. Joint model fully correlated. The event predictor has the same structure as *JM1*. However, the assumption over the random effects in the longitudinal outcomes is

$$\begin{aligned}
\eta_{Yi,s}^{(1)} &= \beta_{01}^{(1)} + \beta_{11}^{(1)} \cdot s + U_{0i}^{(1)} + U_{1i}^{(1)} \cdot s \\
\eta_{Yi,s}^{(2)} &= \beta_{01}^{(2)} + U_{0i}^{(2)} + U_{1i}^{(2)} \cdot s \\
(U_{0i}^{(1)}, U_{1i}^{(1)}, U_{0i}^{(2)}, U_{1i}^{(2)})^\intercal &\sim N_4(\mathbf{0}, Q_{\boldsymbol{U}}^{-1}),
\end{aligned}$$

with $Q_{\boldsymbol{U}}$ a dense matrix. We observe that the cumulative ratio has a linear trend (see Figure 4.4), which explains the additional fixed effect term $\beta_{11}^{(1)} \cdot s$ in comparison with the internal score. Moreover, as we mention in Section 4.2, there is flexibility in how we link the event and the longitudinal processes. We find that linking them only through the random effects provides good performance, but this is not a restriction of this general approach (see Hickey et al., 2016).

We show in Section 4.2.4 that the performance metrics depend on the pair of evaluation times we choose (c and $c+\Delta c$ denoted above). To make the comparison less arbitrary, we evaluate the full range of available starting points ($c = 12, \ldots, 28$) with a fixed time window of 12 months ($\Delta c = 12$), commonly used in the industry. Note that the starting point could have been $c = 1$, but the *Cox_Lag* model limits the comparison due to the lagged observations.

In each validation fold we calculate the $\widehat{AUC}$ and $\widehat{EPE}$ metrics presented in Section 4.2.4 for all the pairs of evaluation times $\{(c, c+12) : c = 12, \ldots, 28\}$. Then, we summarise the metrics for the different pairs of times by calculating

$C_{AUC}^{\Delta c=12}$ and $C_{EPE}^{\Delta c=12}$ (Equations 4.8 and 4.9, respectively). Table 4.2 shows these metrics where each row is a different fold, and the best performance out of the four models is marked in bold. The last row is the average among the ten folds (Avg). First, we observe that, in general terms, the multivariate joint models outperform survival models in both discrimination and calibration metrics. Second, whenever one of the survival models predicts more accurately than the joint models, the difference in the metrics does not seem to be as significant as when we have the opposite.

	Cox		Cox_Lag		JM1		JM2	
Fold	C_{AUC}^{12}	C_{EPE}^{12}	C_{AUC}^{12}	C_{EPE}^{12}	C_{AUC}^{12}	C_{EPE}^{12}	C_{AUC}^{12}	C_{EPE}^{12}
1	**0.6081**	0.5111	0.5874	0.3139	0.5853	**0.3119**	0.5869	**0.3119**
2	**0.6352**	0.4850	0.6240	0.4422	0.6125	**0.4047**	0.6135	0.4048
3	0.5504	0.4766	0.5499	0.3215	**0.6053**	**0.3155**	0.6051	0.3156
4	0.6156	0.4982	**0.6301**	0.3887	0.6228	**0.3672**	0.6191	0.3675
5	**0.5523**	0.5066	0.5170	0.3973	0.5416	**0.3689**	0.5408	0.3691
6	0.6136	0.7152	0.6741	**0.3044**	0.6996	0.3089	**0.7009**	0.3084
7	0.6459	0.4635	0.5850	**0.2713**	0.6716	0.2825	**0.6764**	0.2835
8	0.5944	0.5168	0.5976	0.3105	0.5995	0.3094	**0.6016**	**0.3092**
9	0.7076	0.4832	0.6890	0.3169	0.7702	0.3047	**0.7703**	**0.3046**
10	0.5721	0.5494	0.5653	0.3410	0.6178	0.3268	**0.6195**	**0.3263**
Avg	0.6095	0.5206	0.6019	0.3408	0.6326	**0.3301**	**0.6334**	**0.3301**

Table 4.2: Comparison of the discrimination (C_{AUC}^{12}) and calibration (C_{EPE}^{12}) metrics between the four models for a prediction window of 12 months. Each fold number represents the validation fold in the cross-validation analysis. The last row is the average (Avg) among the ten folds, and the bold number is the best performance metric within each validation fold.

We perform a Bayesian correlated t-test (Benavoli et al., 2017) for both metrics (C_{AUC}^{12} and C_{EPE}^{12}) in order to test the statistical validity of the differences shown in Table 4.2. The test is correlated because the metrics in each fold are not independent since we have overlapping training sets (Nadeau and Bengio, 2000).

Following the recent work Gunnarsson et al. (2021), we also consider two classifiers as practically equivalent when the mean difference of the metric is less than 0.01 and define the *Region of Practical Equivalence* (ROPE) as the interval $[-0.01, 0.01]$. In Figures 4.6 and 4.7 we show the results for all combinations of model pairs for discrimination and calibration, respectively. On the left side of the figures are the reference models (A), and on top are the models we are com-

paring to (B). For instance, we estimate that the *Cox* model has a probability of 0.44 of being better in terms of discrimination than the *Cox_Lag* model, a probability of 0.43 of being equivalent and only 0.13 of being worse. We also observe that the joint models, in comparison to both survival models, are superior, and there is no difference between the two joint models (ROPE-probability 1).

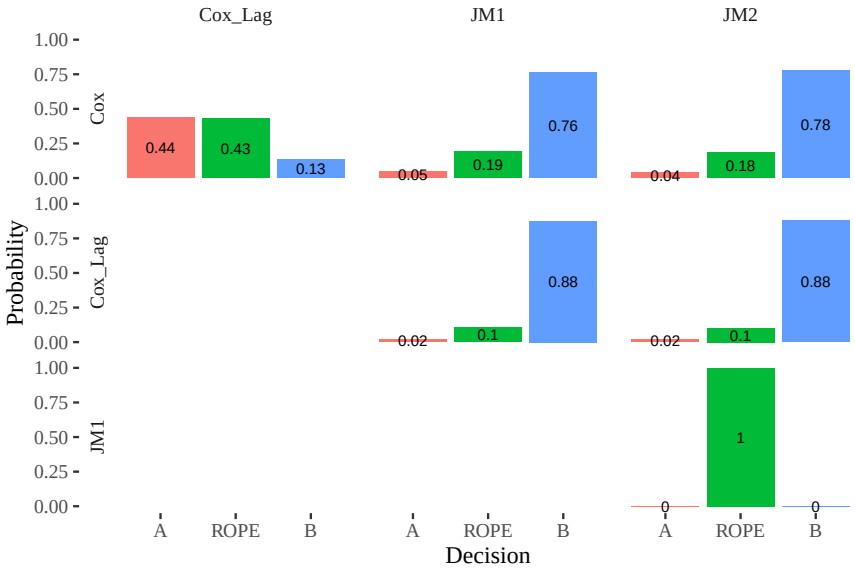

Figure 4.6: Bayesian correlated t-test for the discrimination metric (C^{12}_{AUC}). It shows a three-by-three matrix of bar plots, where each plot compares the reference model named in row (A) and the model we are comparing to in column (B). The bars represent the posterior probabilities of the three possible decisions: A better than B (left bar in red), A practically equivalent to B (centre bar in green) and B better than A (right bar in blue).

In terms of calibration (Figure 4.7), we see that the *Cox* model performs poorly with respect to the three other models. When we compare the *Cox_Lag* model against the joint models, we observe that the probability that these models have the same calibration metrics is 0.45 and a probability of 0.54 in favour of the joint models. Moreover, since we deliberately estimate the *Cox_Lag* model for predicting in a 12-month window (unlike the joint models), we expect it to do well in calibration. Still, the evidence suggests that this model is not better than the joint models (probability of 0.01). Finally, we see no difference between the two joint models (ROPE-probability 1).

So far, we have seen that the benchmarks cannot outperform the joint models for a 12-month forecast horizon ($\Delta c = 12$) for different starting months (c). Still, studying how these models behave when we vary the forecast horizon is also interesting. For a fixed c we obtain the prediction for $[c+1, c+\Delta c]$ for different

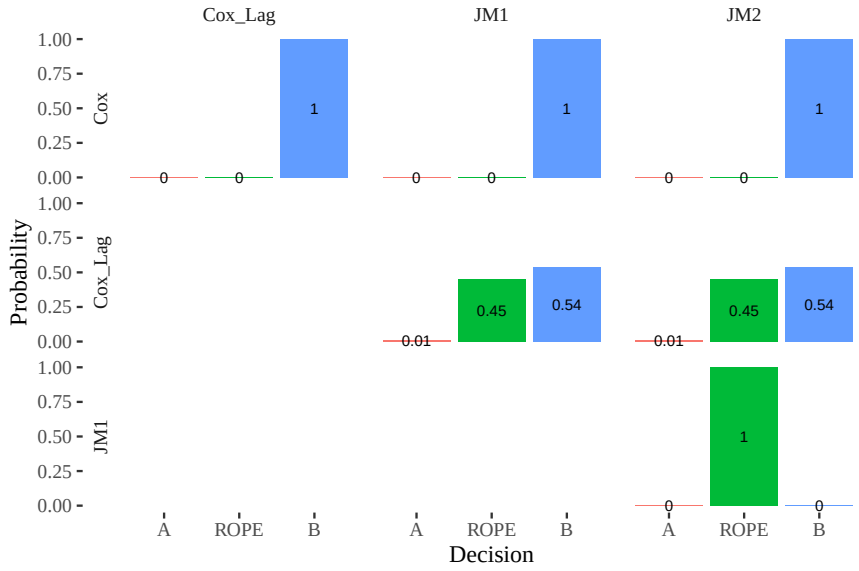

Figure 4.7: Bayesian correlated t-test for the calibration metric (C^{12}_{EPE}). It shows a three-by-three matrix of bar plots, where each plot compares the reference model named in row (A) and the model we are comparing to in column (B). The bars represent the posterior probabilities of the three possible decisions: A better than B (left bar in red), A practically equivalent to B (centre bar in green) and B better than A (right bar in blue).

values of Δc. Figure 4.8 shows the average difference in the $\widehat{AUC}$ with respect to the *Cox* model for Δc ranging from 3 to 24 months and $c = 12$, which is the first period the *Cox_Lag* model can predict. In general, we observe that both joint models have better discrimination for all time windows, a difference that is even stronger for longer horizons.

Analogously to Figure 4.8, in Figure 4.9 we show the average difference in the $\widehat{EPE}$ with respect to the *Cox_Lag* model[4]. We observe that both joint models have practically the same calibration metric (overlapping lines) and that for almost all horizons, this value is below (better) than the *Cox_Lag* level, especially for longer horizons. It is not surprising that for $\Delta c = 12$, the calibration of the reference model is better than the joint models since it was estimated for this Δc. Still, this only includes $c = 12$, and we have shown above that this result does not generalise when considering all c.

We next perform out-of-time data analysis to study the robustness of these results. In practice, models are applied to new data at later periods than the ones used in the construction stage. This, for example, happens when a bank is interested

[4]We discard the *Cox* model from this plot because its performance is considerably inferior concerning the others.

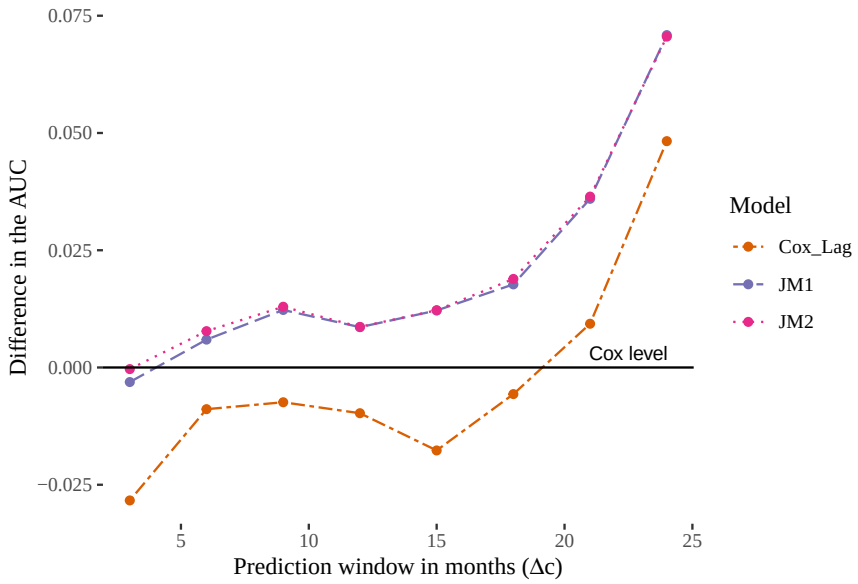

Figure 4.8: Average difference in the $\widehat{AUC}$ with respect to the *Cox* model, for fixed $c = 12$ and variable Δc.

in classifying new customers. We study and compare how these models perform in an out-of-time scenario following a similar analysis to the previous one. We could estimate a model per specification using all the training data, but since we have already estimated ten models per specification, we use each to calculate the out-of-time performance.

In Table 4.3 we show the results for each model. We note that none of the traditional survival approaches can outperform the joint models, a result further supported by the Bayesian correlated t-test shown in Figures 4.10 and 4.11 for discrimination and calibration, respectively.

From these figures, we note that the survival models are practically equivalent in terms of discrimination (ROPE-probability 1), but the *Cox_Lag* model outperforms the *Cox* in terms of calibration. Moreover, both joint models have a probability of 1 of being better than the survival models for both metrics, and there is not much difference between them (ROPE-probability 1).

The discrimination and calibration performances for different time windows Δc are shown in Figures 4.12 and 4.13, respectively. We observe that both joint models have better $\widehat{AUC}$ than the survival models for basically all the horizons. In calibration, we now see that for all the Δc, the $\widehat{EPE}$ for each joint model is lower than for the *Cox_Lag* model. Also, the minimum difference is again

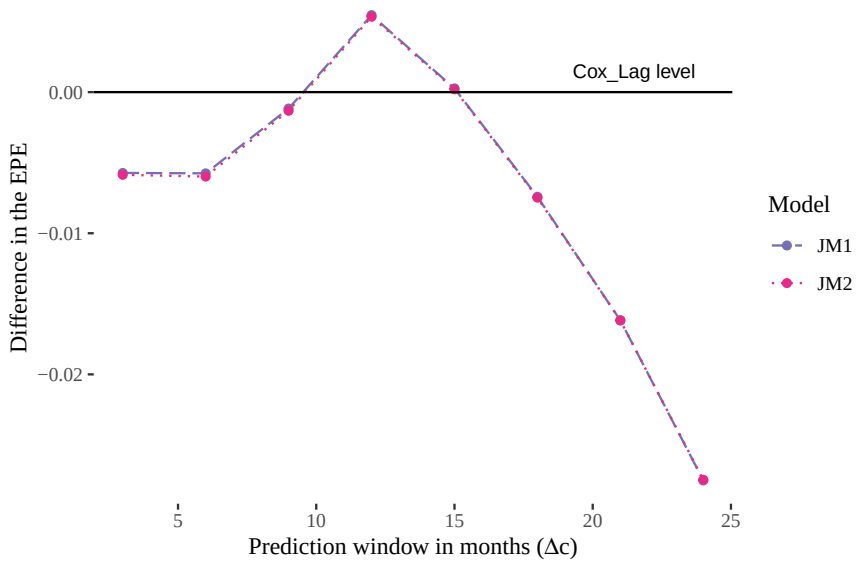

Figure 4.9: Average difference in the $\widehat{EPE}$ with respect to the *Cox_Lag* model, for fixed $c = 12$ and variable Δc.

obtained at $\Delta c = 12$, and it increases for longer horizons.

4.5 Discussion

The joint modelling approach applied in the credit context is appealing compared to traditional survival analysis. It allows us to incorporate potentially endogenous TVCs and provide a dynamic prediction framework that correctly updates once new information is collected. However, joint model estimations are computationally intensive when maximum likelihood or MCMC schemes are used, which is even more critical in large datasets with more than one endogenous TVC (multivariate). That is commonly the case with credit-related applications.

In this chapter, we make two methodological and two empirical contributions. First, we propose a fast and accurate joint model of bivariate longitudinal outcomes and discrete survival data based on the INLA framework. We study this model via simulation analysis. Second, we introduce a methodology for individual survival predictions using the Laplace method that leads to more accurate approximations than comparable approaches. From the empirical level, first, we present a multivariate joint model in the credit risk literature, specifically for predicting the probability of full prepayment in a consumer loan portfolio. Second, we show that for this particular application, the multivariate joint models out-

	Cox		Cox_Lag		JM1		JM2	
Fold	C^{12}_{AUC}	C^{12}_{EPE}	C^{12}_{AUC}	C^{12}_{EPE}	C^{12}_{AUC}	C^{12}_{EPE}	C^{12}_{AUC}	C^{12}_{EPE}
1	0.5531	0.5030	0.5520	0.4001	**0.5668**	**0.3755**	0.5664	0.3761
2	0.5531	0.4573	0.5523	0.4025	**0.5647**	**0.3761**	**0.5647**	0.3766
3	0.5533	0.4639	0.5525	0.4118	**0.5640**	**0.3749**	0.5629	0.3755
4	0.5532	0.4756	0.5521	0.4079	0.5676	0.3763	**0.5689**	**0.3755**
5	0.5549	0.4917	0.5523	0.4082	0.5707	**0.3749**	**0.5710**	0.3751
6	0.5509	0.6729	0.5520	0.4005	**0.5704**	0.3770	0.5691	**0.3767**
7	0.5534	0.4816	0.5513	0.4016	**0.5662**	0.3759	0.5629	**0.3754**
8	0.5519	0.5097	0.5520	0.4069	0.5692	0.3766	**0.5694**	**0.3761**
9	0.5506	0.4970	0.5527	0.4020	**0.5665**	**0.3752**	0.5658	0.3754
10	0.5532	0.5274	0.5522	0.4293	**0.5701**	**0.3762**	0.5687	0.3765
Avg	0.5528	0.5080	0.5521	0.4071	**0.5676**	**0.3759**	0.5670	**0.3759**

Table 4.3: Comparison of the discrimination (C^{12}_{AUC}) and calibration (C^{12}_{EPE}) metrics between the four models for a prediction window of 12 months. Each fold number represents the hold-out fold when training the model. The predictions are made in the out-of-time dataset. The last row is the average (Avg) among columns, and the bold number is the best performance metric per row.

perform standard survival approaches in out-of-sample and out-of-time analyses.

As a new approach to credit risk modelling, many possibilities for future development remain that we believe could further enhance its use. For example, exploring the idea of including TVCs related to other credit products, such as credit card transactions, and studying how that relates to the event of interest. Conceptually, this is straightforward since we are not restricted to simultaneously collecting the survival and longitudinal data. Although including a larger number of measurements increases the computational cost, we believe the INLA approach we propose here could be a viable path. Moreover, we could also study new link structures between the event and longitudinal processes where the effect among them changes depending on the stage of the credit. The joint models' approach offers the flexibility to explore these and other compelling topics.

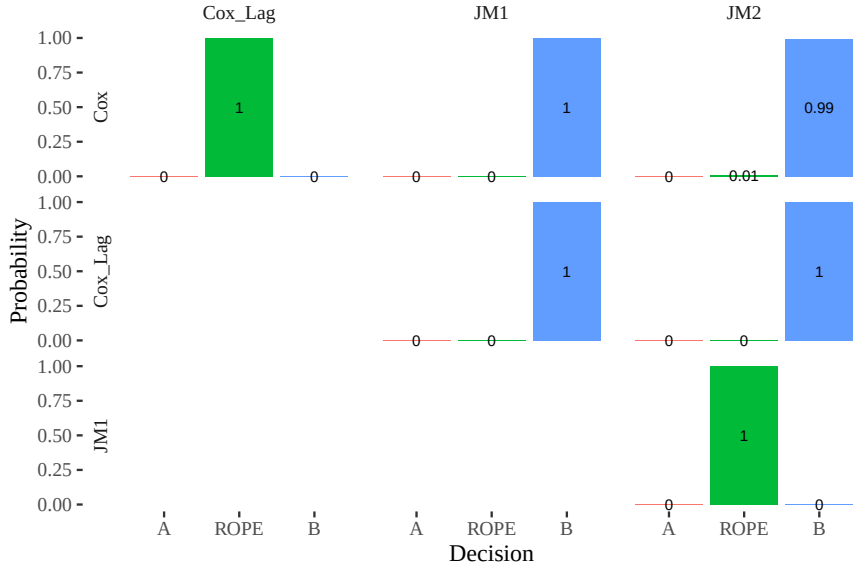

Figure 4.10: Bayesian correlated t-test for the discrimination metric (C^{12}_{AUC}) shown as in Figure 4.6 and applied to the out-of-time dataset.

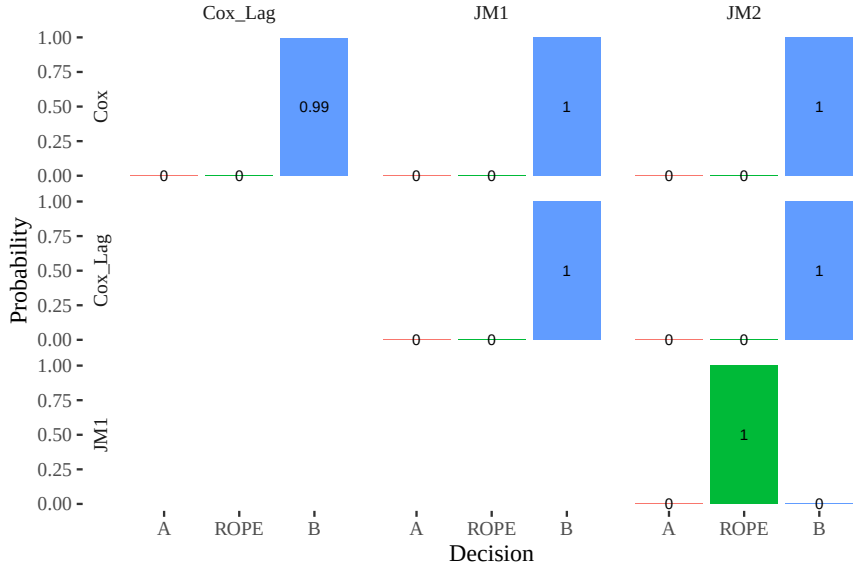

Figure 4.11: Bayesian correlated t-test for the calibration metric (C^{12}_{EPE}) shown as in Figure 4.7 and applied to the out-of-time dataset.

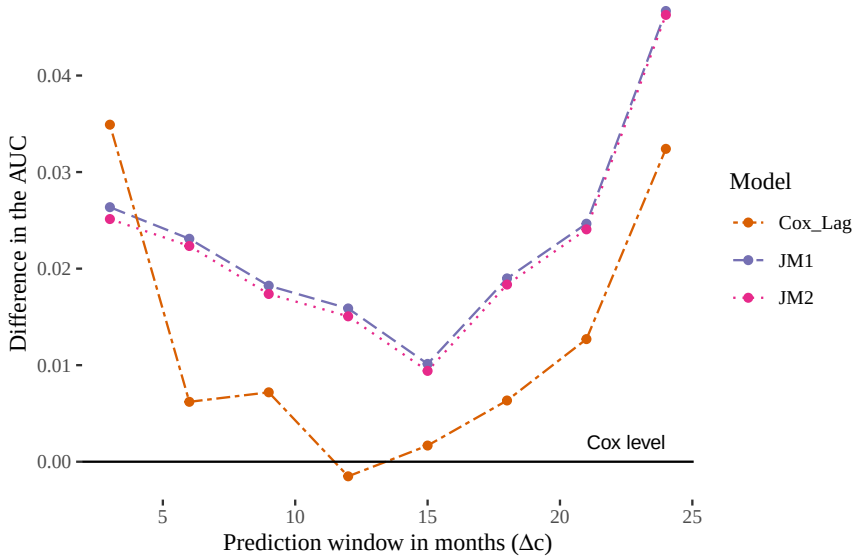

Figure 4.12: Average difference in the $\widehat{AUC}$ with respect to the *Cox* model, for fixed $c = 12$ and variable Δc. Results from the out-of-time analysis.

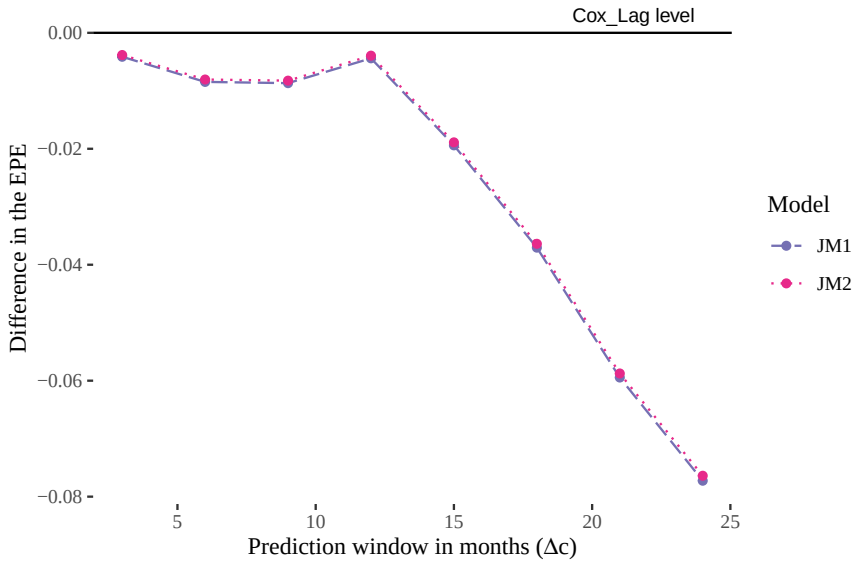

Figure 4.13: Average difference in the $\widehat{EPE}$ with respect to the *Cox_Lag* model, for fixed $c = 12$ and variable Δc. Results from the out-of-time analysis.

Chapter 5

Spatio-Temporal Joint Models

In Chapters 3 and 4, we have studied the joint model approach with autoregressive terms in the longitudinal sub-model and included more than one time-varying covariate. Generally, both chapters introduce innovations more oriented towards the longitudinal sub-model(s). In this chapter, however, we explore new ways of improving the survival sub-model by incorporating spatial and spatio-temporal effects in its predictor. The aim is to capture the survival effect due to the evolution of the unobserved heterogeneity among borrowers in the same region.

The chapter is organised as follows. In Section 5.1, we introduce the relevant literature on spatial and spatio-temporal effects to situate our contributions in both the credit risk and the joint model context. In Section 5.2, we present the spatio-temporal joint model (STJM) and describe how its estimation is performed. Moreover, for model comparison purposes, we also introduce a new implementation of the *cross-validated Dynamic Conditional Likelihood* (cvDCL Rizopoulos et al., 2016) that can be nicely incorporated into our estimation framework without requiring extensive additional calculations. In Section 5.3, we build joint models to predict the time to full prepayment event on a US mortgage portfolio and study how the inclusion of spatial and spatio-temporal effects can improve the models' performance. In Section 5.4 we conclude and comment on further research.

5.1 Introduction

The modelling of the survival time generally includes time-fixed and time-varying covariates (TVCs). Along with this thesis, we have seen two main appealing features of the joint modelling approach compared to the standard survival credit risk model. Namely, when the TVCs are endogenous, joint models offer a sound statistical procedure to handle the mutual evolution of the survival process and the endogenous TVCs. In addition, by jointly modelling both the survival and the TVCs, we encounter a natural prediction framework that does not rely on lagged values or exogeneity assumptions as commonly done otherwise.

In previous chapters, we exploit joint models' flexibility for modelling endogenous TVCs. Here, we are instead interested in finding more flexible representations of the survival predictor that may help us to explain the credit data better. Additionally, it is increasingly common to incorporate geographical information about the borrower into the databases (Goodstein et al., 2017; Gupta, 2019; Calabrese and Crook, 2020), giving way to models that also account for spatial clustering and its variation in time. To this end, we propose a Bayesian hierarchical joint model in discrete time that includes spatial and spatio-temporal effects in the baseline hazard and aims to predict full prepayment events in US mortgages. This approach captures the survival impact due to the evolution of the unobserved spatial confounders among borrowers and lets us leverage information across neighbouring areas.

A few studies in the realm of mortgage credit analysis include spatial dependency. For example, Goodstein et al. (2017) by analysing a large mortgage dataset, and after controlling for known default factors, establish the impact of the surrounding areas in strategic mortgage default. That is when the borrower chooses to default because the economic benefits of doing so outweigh its costs (unlike borrowers who default because they have no other choice). In the same line, Guiso et al. (2013); Towe and Lawley (2013) encounter strong evidence that social interactions among neighbours influence the propensity of strategic default.

The spatial contagion in mortgage default has been acknowledged due to different causes. For example, the reduction of the property's value in the neighbourhood can increase the default propensity (particularly those by choice). That property's value can be affected by the neighbourhood characteristics, such as increasing

crime rate, vandalisation, etc., or even legislative reasons (Pence, 2006). On the other hand, due to the increased defaults, banks might also limit the credit options in those areas (e.g. renegotiation), deepening even more, the correlated effect.

However, spatial contagion in credit risk has not been limited only to mortgages. Calabrese et al. (2019), for example, include spatial dimensions to predict credit default on SMEs in the UK. Medina-Olivares et al. (2022b) find that spatial dependency can improve the performance of credit scoring models for microfinance in China. Moreover, spatial contagion can also be found in events other than defaults. Gupta (2019) finds that early repayment activity in mortgage loans has a significant spatial dependence that might be related to similar reasons as the ones associated with the default. The author notes that from a borrower-driven point of view, a decrease in property values can also decrease borrowers' propensity to seek new refinancing alternatives. Additionally, from a lender-driven perspective, credit extensions or renegotiation can be reduced if banks estimate a drop in property prices for some locations or other foreclosure externalities.

In the context of survival models with applications to credit risk, Calabrese and Crook (2020) claim to be the first paper to present a survival model with spatial contagion. They incorporate time and spatial-varying coefficients in a survival model that predicts time to default in UK mortgage loans, showing better accuracy than relevant benchmarks. However, they overlook possible endogeneity in the TVCs included in the model. And they do not present a prediction framework that accounts for the future paths of these TVCs as the joint model approach does.

As far as the joint model literature is concerned, Zhou et al. (2008) propose a joint model in continuous time for modelling two linked time-to-event outcomes, assuming a Weibull baseline distribution with spatially correlated frailties. Moreover, the work of Ratcliffe et al. (2004) incorporates spatial clustering as univariate independent random effects. More in the spirit of this work, Martins et al. (2016) propose a joint model with spatial random effects for analysing AIDS data in Brazil. They assume an intrinsic conditional autoregressive model (ICAR, see Besag et al., 1991) as a prior distribution for the unobserved spatial effects, as we also do in this work.

We make four contributions to the literature. First, we introduce a discrete-time joint model with a flexible baseline hazard that can handle spatial and spatio-temporal interactions. We denote this model Spatio-Temporal Joint Model (STJM). Second, to estimate this model in an extensive mortgage loans dataset, we implement it using the INLA methodology (Rue et al., 2009). That lets us scale the model to a dataset with a total of 2,559,056 observations, and as far as we know, it is the largest one in the context of joint models at the time of writing. Third, to compare different model specifications, we propose a new implementation of the *cross-validated Dynamic Conditional Likelihood* (cvDCL, see Rizopoulos et al., 2016)[1] that uses already calculated quantities by the INLA methodology, making its estimation computational convenient. And forth, we apply the STJM to predict full prepayment events in US mortgage loans. We show that the inclusion of the spatial components can consistently improve the performance of the joint model for all the evaluation times considered. However, we also found in our empirical analysis that the performance improvements are less conclusive when including spatio-temporal effects on top of the spatial main effects.

5.2 Methodology

5.2.1 Spatio-Temporal Joint Model (STJM)

Consider a total of N mortgage loans where the properties are distributed over A areas. Each area $a = 1, \ldots, A$ has a total of N_a properties, i.e. $\sum_{a=1}^{A} N_a = N$. For each mortgage loan i ($i = 1, 2, \ldots, N$), the following characteristics are known: the location $a_i \in \{1, \ldots, A\}$, the date when the loan is originated t_i^0, the event indicator δ_i that takes the value of 1 if the full prepayment occurs and 0 otherwise, and the time elapsed from origination of the loan to the last available observation $t_i \leq T$, where T corresponds to the duration of the study. We assume that at time t_i either the full prepayment happens ($\delta_i = 1$), or the observation is right-censored ($\delta_i = 0$). We are also provided with a vector of time-fixed covariates $\boldsymbol{z}_i$, and a loan-specific covariate collected at multiple points in time on a regular basis (in our case, on a monthly basis), denoted by $y_{i,s}$ for $s = 1, \ldots, t_i$. This TVC

[1] The cvDCL is a cross-entropy estimate of the cross-validatory posterior predictive conditional density.

corresponds to our joint model framework' longitudinal outcome. As used in the previous chapters, we distinguish the realisations of a random variable in lower case.

We aim to understand the relationship of these data in jointly modelling the time to event T_i and the longitudinal outcome $Y_{i,s}$ up to a given endpoint for the i-th loan associated with area a_i. In the following, we describe the proposed approach for the longitudinal and survival processes.

Longitudinal process

Following the notation introduced in Chapter 2, assume the longitudinal outcome $Y_{i,s}$ is modelled by a mixed-effect model (Laird and Ware, 1982) where the predictor $\eta_{Yi,s}$ is structured by fixed effects, $\boldsymbol{q}_{i,s}^{\intercal}\boldsymbol{\beta}_1$, and random effects, $\boldsymbol{d}_{i,s}^{\intercal}\boldsymbol{U}_i$. $\boldsymbol{\beta}_1$ is the vector of coefficients related to the covariates $\boldsymbol{q}_{i,s}$ and $\boldsymbol{d}_{i,s}$ is the design vector related to the random effects $\boldsymbol{U}_i$ of dimension r. Specifically,

$$\begin{aligned}(Y_{i,s}|\eta_{Yi,s},\tau_Y) &\sim N(\eta_{Yi,s},\tau_Y^{-1})\\ \eta_{Yi,s} &= \boldsymbol{q}_{i,s}^{\intercal}\boldsymbol{\beta}_1 + \boldsymbol{d}_{i,s}^{\intercal}\boldsymbol{U}_i\\ \boldsymbol{U}_i|Q_{\boldsymbol{U}} &\sim N_r(\boldsymbol{0},Q_{\boldsymbol{U}}^{-1}),\end{aligned} \tag{5.1}$$

where τ_Y is the precision parameter of the innovations. We assume that $\boldsymbol{U}_i$ are mutually independent among mortgage loans and distributed as a zero-mean multivariate Gaussian distribution with $r \times r$ precision matrix $Q_{\boldsymbol{U}}$. We also assume that observations within each loan are conditionally independent given the random effects. Therefore, the random effects account for the correlation between these different observations.

Survival process

Following the discrete-time survival formulation used in Chapter 2, we represent the random variable T_i as the sequence of binary random variables $X_{i,s}$ that takes the value 1 if the loan i is fully prepaid at time $s = t_i$ and 0 otherwise. Furthermore, we relate $X_{i,s}$ with the predictor $\eta_{Xi,s}$ through a logit link function as follows

$$\begin{aligned}(X_{i,s}|X_{i,s-1}=0,\eta_{Xi,s}) &\sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{Xi,s}))\\ \eta_{Xi,s} &= \nu_{a_i,s} + \boldsymbol{z}_i^{\intercal}\boldsymbol{\beta}_2 + \lambda(\boldsymbol{d}_{i,s}^{\intercal}\boldsymbol{U}_i),\end{aligned} \tag{5.2}$$

where $\nu_{a_i,s}$ is the baseline risk which varies with both time and space and extends the entire discrete domain of $a_i \in \{1,\dots,A\}$ and $s \in \{1,\dots,T\}$. Moreover, $\boldsymbol{\beta}_2$ is the vector of coefficients related to covariates $\boldsymbol{z}_i$ and λ is the association parameter between the survival time and the random effects $\boldsymbol{d}_{i,s}^{\intercal}\boldsymbol{U}_i$. Therefore, the random effects play an important role in both the longitudinal and survival processes. In the longitudinal process (Equation 5.1), account for the correlation between repeated measurements and, in the survival process (Equation 5.2), together with λ, account for the degree of association with the longitudinal outcome.

Following Chang et al. (2013), which decompose the spatio-temporal effects additively for a survival model, we consider for our joint model $\nu_{a,s} = \nu_0 + v_s + u_a + \delta_{a,s}$, where ν_0 is the overall average, v_s is the temporal main effect, u_a is the spatial main effect and $\delta_{a,s}$ is the spatio-temporal interaction. In the following, we describe the terms v_s, u_a and $\delta_{a,s}$.

Temporal main effects (v_s): Let us denote the vector of temporal effects as $\boldsymbol{v} = (v_1,\dots,v_T)^{\intercal}$. As in Chapter 4, we assume that these effects are represented by a second-order random walk model (see Lindgren and Rue, 2008) which has the following joint density

$$\begin{aligned} \boldsymbol{v}|\tau_v &\propto \exp\left(-\frac{\tau_v}{2}\sum_{s=3}^{T}(v_s - 2v_{s-1} + v_{s-2})^2\right) \\ &= \exp\left(-\frac{\tau_v}{2}\boldsymbol{v}^{\intercal}R_v\boldsymbol{v}\right), \end{aligned} \tag{5.3}$$

where τ_v is a precision parameter and the $T \times T$ matrix R_v is the so-called *structure matrix* (Rue and Held, 2005) defined as (the zeros are not shown)

$$R_v = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}.$$

Spatial main effects (u_a): For the spatial effects $\boldsymbol{u} = (u_1,\dots,u_A)^{\intercal}$ we assume an intrinsic conditional autoregressive model (ICAR, see Besag et al., 1991) which accounts for the fact that close areas might have similar repayment behaviour

(see, for example, Calabrese and Crook (2020) for default prediction). The joint density for the ICAR model is described as

$$\boldsymbol{u}|\tau_u \propto \exp\left(-\frac{\tau_u}{2}\sum_{a\sim a'}(u_a - u_{a'})^2\right), \tag{5.4}$$

where τ_u is a precision parameter and $a \sim a'$ denotes that the two areas are neighbours. There are many ways of defining a "neighbour" and ultimately will depend on the application (Freni-Sterrantino et al., 2018). Here, we consider the standard definition where two areas are regarded as neighbours if they share a common border. Another notion of two connected areas, for example, can be defined in terms of the distance between the centroids of these areas (e.g. Goodstein et al., 2017; Medina-Olivares et al., 2022b). The exploration of different ways to define neighbours, however, is beyond the scope of this chapter and the reader is referred to Banerjee et al. (2014, Ch. 4) for further discussion on this topic.

For this specification, the corresponding elements of the $A \times A$ structure matrix R_u of Equation 5.4 are

$$(R_u)_{aa'} = \begin{cases} m_a & a = a' \\ -1 & a \sim a' \\ 0 & \text{otherwise,} \end{cases}$$

where m_a is the number of neighbours of area a. The interpretation of the ICAR model is made more accessible from the full conditional density given by

$$(u_a|\boldsymbol{u}_{-a}, \tau_u) \sim N\left(\frac{1}{m_a}\sum_{a':a\sim a'} u_{a'}, \frac{1}{\tau_u m_a}\right),$$

where $\boldsymbol{u}_{-a}$ represents the set of spatial effects without the area a. Hence, u_a has a local mean of $\sum_{a':a\sim a'} u_{a'}/m_a$, which is the average value of the spatial effects from the neighbours, and a variance that is inversely related to the number of neighbours m_a. That means the more neighbours there are, the more certainty there is of the effect.

Spatio-temporal interactions ($\delta_{a,s}$): For the spatio-temporal interactions $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{A1}, \ldots, \delta_{1T}, \ldots, \delta_{AT})^\intercal$ we follow the approach from Clayton (1996), and further detailed in Knorr-Held (2000), in which the structure matrix R_δ can

be derived as the Kronecker product of the structure matrices from the temporal and spatial main effects, i.e. $R_\delta = R_v \otimes R_u$. Then, the corresponding joint density is (Schrödle and Held, 2011)

$$\boldsymbol{\delta}|\tau_\delta \propto \exp\Bigg(-\frac{\tau_\delta}{2} \sum_{s=3}^{T} \sum_{a\sim a'} \Big[(\delta_{a,s} - 2\delta_{a,s-1} + \delta_{a,s-2}) - (\delta_{a',s-2} - 2\delta_{a',s-1} + \delta_{a',s}) \Big]^2 \Bigg), \quad (5.5)$$

where τ_δ is the corresponding precision parameter.

In the spatial literature, it is well-known that structured additive predictors formed by Equations 5.3, 5.4 and 5.5 lead to identifiability problems (see, e.g. Knorr-Held, 2000; Goicoa et al., 2018). Therefore, we must set constraints over the random effects $\boldsymbol{v}$, $\boldsymbol{u}$ and $\boldsymbol{\delta}$. To get appropriate identifiability constraints, we follow Goicoa et al. (2018) who use reparametrisations over the structure matrices R_v, R_u and R_δ using spectral decomposition. These reparametrisations conduct to the following constraints: $\sum_{s=1}^{T} v_s = 0$, $\sum_{a=1}^{A} u_a = 0$, $\sum_{s=1}^{T} \delta_{a,s} = 0$ for $a = 1, \ldots, A$ and $\sum_{a=1}^{A} \delta_{a,s} = 0$ for $s = 1, \ldots, T$.

5.2.2 Estimation

From Section 5.2.1 we know that the random effects $\boldsymbol{U}_i$ are shared between the longitudinal and the survival processes. We have also seen in Chapter 2 that the main assumption in the joint model approach is that these two processes are conditionally independent given the random effects (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Tsiatis and Davidian, 2004). Therefore, the joint distribution of the observation variables $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,t_i})^\intercal$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i,t_i})^\intercal$ for loan i conditional on the random effects is

$$p(\boldsymbol{y}_i, \boldsymbol{x}_i|\boldsymbol{U}_i, \Theta) = \prod_{s=1}^{t_i} p(y_{i,s}|\boldsymbol{U}_i, \Theta) p(x_{i,s}|\boldsymbol{U}_i, \Theta), \quad (5.6)$$

where Θ represents the vector of parameters included in both processes. It follows from Equation 5.1 that

$$\begin{aligned} p(y_{i,s}|\boldsymbol{U}_i, \Theta) &= \left(\frac{\tau_Y}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau_Y (y_{i,s} - \eta_{Yi,s})^2}{2} \right) \\ &= \left(\frac{\tau_Y}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau_Y (y_{i,s} - \boldsymbol{q}_{i,s}^\intercal \boldsymbol{\beta}_1 - \boldsymbol{d}_{i,s}^\intercal \boldsymbol{U}_i)^2}{2} \right), \end{aligned}$$

and from Equation 5.2

$$\begin{aligned} p(x_{i,s}|\boldsymbol{U}_i,\Theta) &= [\text{logit}^{-1}(\eta_{Xi,s})]^{x_{i,s}}[1-\text{logit}^{-1}(\eta_{Xi,s})]^{1-x_{i,s}} \\ &= [\text{logit}^{-1}(\nu_{a_i,s}+\boldsymbol{z}_i^\intercal\boldsymbol{\beta}_2+\lambda(\boldsymbol{d}_{i,s}^\intercal\boldsymbol{U}_i))]^{x_{i,s}} \\ &\quad\times[1-\text{logit}^{-1}(\nu_{a_i,s}+\boldsymbol{z}_i^\intercal\boldsymbol{\beta}_2+\lambda(\boldsymbol{d}_{i,s}^\intercal\boldsymbol{U}_i))]^{1-x_{i,s}}. \end{aligned}$$

Hence, the contribution of the i-th loan to the observation density is

$$\begin{aligned} p(\boldsymbol{y}_i,\boldsymbol{x}_i|\Theta) &= \int p(\boldsymbol{y}_i,\boldsymbol{x}_i|\boldsymbol{U}_i,\Theta)p(\boldsymbol{U}_i|\Theta)\text{d}\boldsymbol{U}_i \\ &= \int\prod_{s=1}^{t_i} p(y_{i,s}|\boldsymbol{U}_i,\Theta)p(x_{i,s}|\boldsymbol{U}_i,\Theta)p(\boldsymbol{U}_i|\Theta)\text{d}\boldsymbol{U}_i, \end{aligned} \tag{5.7}$$

where $p(\boldsymbol{U}_i|\Theta)$ is as zero-mean multivariate Gaussian with precision matrix $Q_{\boldsymbol{U}}$ (Section 5.2.1), i.e. $p(\boldsymbol{U}_i|\Theta) = (2\pi)^{-r/2}|Q_{\boldsymbol{U}}|^{1/2}\exp\left(-\boldsymbol{U}_i^\intercal Q_{\boldsymbol{U}}\boldsymbol{U}_i/2\right)$.

Denote the complete set of observation variables as $\mathcal{D} = \{\boldsymbol{y}_i,\boldsymbol{x}_i : i = 1,\ldots,N\}$. The joint posterior distribution follows $p(\Theta|\mathcal{D}) \propto p(\mathcal{D}|\Theta)p(\Theta)$, where $p(\mathcal{D}|\Theta) = \prod_i^N p(\boldsymbol{y}_i,\boldsymbol{x}_i|\Theta)$ is the overall observation density and $p(\Theta)$ the joint prior.

We could theoretically estimate this model specification with simulation-based schemes as done in Chapter 3 for the joint model with autoregressive terms. As we noted, this strategy is computationally expensive, or it might be even infeasible for applications with big datasets. Furthermore, and in line with the estimation strategy followed in Chapter 4, we propose to use the INLA methodology (Rue et al., 2009). As mentioned earlier, INLA provides accurate estimations of the posterior at a lower computational cost and is easily accessible through the `R-INLA` software package for R (`https://www.r-inla.org/`). This methodology applies to models belonging to the class of latent Gaussian models (LGM), a flexible and widely used class of models. For example, most structured Bayesian additive models are of this type (see Fahrmeir and Tutz, 1994; Gelman et al., 2013). The STJM is also found in this class of models, as shown next.

Following Sections 2.2.4 and 4.2.2, we identify the latent field $\boldsymbol{\mu} = (\boldsymbol{\eta}_Y,\boldsymbol{\eta}_X,\boldsymbol{U},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\nu_0,\boldsymbol{v},\boldsymbol{u},\boldsymbol{\delta})$ which is the set of unobserved variables in the STJM. The terms $\boldsymbol{\eta}_Y$ and $\boldsymbol{\eta}_X$ correspond to the predictors described in Equations 5.1 and 5.2, each of them with $\sum_i^N t_i$ elements, and because the rest of the elements are latent variables, $\boldsymbol{\mu}$ is referred as a latent field. Moreover, since we assume that $\boldsymbol{\mu}$ follows a zero-mean multivariate Gaussian distribution, $\boldsymbol{\mu}$ is called latent Gaussian field (Rue and Held, 2005).

Concretely, we assume that the coefficients $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ν_0 follows a zero-mean Gaussian distribution with precision matrix $\tau_f \boldsymbol{I}$, with $\boldsymbol{I}$ the identity matrix of the corresponding dimension and τ_f a precision parameter, commonly set as fixed and close to zero in the model (large prior variance). Moreover, as mentioned in Section 5.2.1, $\boldsymbol{U}_i|Q_{\boldsymbol{U}} \sim N(\boldsymbol{0}, Q_{\boldsymbol{U}}^{-1})$ and the terms $\boldsymbol{v}$, $\boldsymbol{u}$ and $\boldsymbol{\delta}$ have priors with Gaussian kernels (see Equations 5.3, 5.4 and 5.5, respectively). Hence, the precision matrix of the latent Gaussian field $\boldsymbol{\mu}$, which encompasses all the individual precision matrices, is denoted as $Q(\boldsymbol{\theta}_1)$, with $\boldsymbol{\theta}_1$ the corresponding set of hyperparameters. In our case $\boldsymbol{\theta}_1 = (\tau_f, Q_{\boldsymbol{U}}, \lambda, \tau_v, \tau_u, \tau_\delta)$. Although the dimension of the matrix $Q(\boldsymbol{\theta}_1)$ can be very large, INLA takes advantage in terms of computation given the sparsity of this matrix (Rue et al., 2009).

Furthermore, denote as $\boldsymbol{\theta}_2$ the set of hyperparameters that have a direct impact on the observation density, which in our case is made up only by the precision parameter τ_Y. We can reformulate Equation 5.6 to the INLA notation as $p(\boldsymbol{y}_i, \boldsymbol{x}_i|\boldsymbol{U}_i, \Theta) = \prod_{s=1}^{t_i} p(\mathcal{D}_{i(s)}|\mu_{i(s)}, \boldsymbol{\theta_2})$, which makes the overall observation density $p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\theta}_2) = \prod_i^N \prod_{s=1}^{t_i} p(\mathcal{D}_{i(s)}|\mu_{i(s)}, \boldsymbol{\theta_2})$ and, thus can be easily written as $p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\theta}_2) = \prod_{j \in \mathcal{J}} p(\mathcal{D}_j|\mu_j, \boldsymbol{\theta_2})$ by changing the corresponding indexes. This last expression shows, as required by INLA methodology, that the observation density is conditional independent (see Section 2.2.5).

Finally, denoting the complete set of hyperparameters as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, we recover the same formulation described in Section 2.2.5. In other words, we see that the STJM belongs to the class of latent Gaussian models, and we can then estimate it with INLA. In particular, we are interested in the posterior marginals, $p(\mu_i|\mathcal{D})$ and $p(\theta_j|\mathcal{D})$, specified by

$$p(\mu_i|\mathcal{D}) = \int p(\mu_i|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}$$
$$p(\theta_j|\mathcal{D}) = \int p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}_{-j}.$$

Refer to Section 2.2.5 to see how INLA methodology estimates these integrals.

5.2.3 Bayesian model selection with INLA

We are interested in selecting the model that best predicts the prepayment event conditional that loan i has not prepaid up to a time point t. To distinguish which model predicts better conditional on the collected observations, we follow the

approach of Rizopoulos et al. (2016) and adapt it to both the INLA estimation procedure and the STJM formulation, as described in the following.

The authors propose to choose the model that minimises the cross-entropy of the cross-validatory posterior predictive conditional density of the survival outcome. Concretely, assume that for model $M_k \in \{M_1, \ldots, M_K\}$ at time t, we are interested in estimating $p(T_i|T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, M_k)$ (also termed cross-validatory posterior predictive conditional density of the survival outcome), where $\boldsymbol{y}_i(t)$ is the set of historical observations for the longitudinal outcome of loan i up to time t, i.e. $\boldsymbol{y}_i(t) = \{y_{i,s} : s \leq t\}$, and $\mathcal{D}_{-i}$ represents the data omitting loan i. The best model $M_{\tilde{k}}$, with $\tilde{k} \in \{1, \ldots, K\}$, is considered the one that minimises the cross-entropy $E(-\log\{p(T_i|T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, M_{\tilde{k}})\})$, where the expectation is taken with respect to the model under which the data have been generated (this model does not need to be part of the K models, as it happens in practice).

To account for the censored cases, Rizopoulos et al. (2016) propose to use the available information and termed this estimate as the *cross-validated Dynamic Conditional Likelihood* (cvDCL) defined as[2]

$$\text{cvDCL}(t) = \frac{1}{N_t} \sum_{i=1}^{N} -I(T_i > t) \log\{p(T_i, \delta_i|T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})\}, \tag{5.8}$$

where N_t is the number of loans at risk at time t, i.e. $N_t = \sum_{i=1}^{N} I(T_i > t)$.

As mentioned in Section 5.2.2, the INLA methodology estimates the posterior marginals of the latent field $\boldsymbol{\mu}$ and the hyperparameters $\boldsymbol{\theta}$. However, once the model is estimated, INLA also allows generating samples from the approximated posterior density. We propose to take advantage of this feature to calculate the expression in Equation 5.8 through Monte Carlo integration, as shown below.

First, note that

$$\frac{p(\boldsymbol{\theta}|T_i, \delta_i, T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}) p(T_i, \delta_i|T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})}{p(T_i, \delta_i|T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta})} = p(\boldsymbol{\theta}|T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}),$$

[2]We omit the explicit notation on model k, but the conditional in the model is implicitly assumed.

and integration of this last expression with respect to $\boldsymbol{\theta}$ leads to

$$\begin{aligned} p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})^{-1} &= \int \frac{p(\boldsymbol{\theta} | T_i, \delta_i, T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \\ &\approx \int \frac{p(\boldsymbol{\theta} | \mathcal{D})}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \\ &\approx \sum_w \frac{\hat{p}(\boldsymbol{\theta}_w | \mathcal{D})}{\hat{p}(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w)} \Delta_w. \end{aligned} \tag{5.9}$$

The integration grid $\{\boldsymbol{\theta}_w, \Delta_w\}$ of $\boldsymbol{\theta}$ is constructed by INLA when estimating the model, where Δ_w represents the integration weights (see Section 2.2.5).

Moreover, note that the denominator $p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w)$ follows

$$\begin{aligned} &\frac{p(\boldsymbol{\mu} | T_i, \delta_i, T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w) p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w)}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{\mu})} \\ &\qquad = p(\boldsymbol{\mu} | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w), \end{aligned}$$

where $\boldsymbol{\mu}$ is the latent field described in Section 5.2.2. Thus, integrating this last expression with respect to $\boldsymbol{\mu}$ gives

$$\begin{aligned} p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w)^{-1} &= \int \frac{p(\boldsymbol{\mu} | T_i, \delta_i, T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w)}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{\mu})} \mathrm{d}\boldsymbol{\mu} \\ &= \int \frac{p(\boldsymbol{U}_i, \boldsymbol{\mu}_{-\boldsymbol{U}_i} | T_i, \delta_i, T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w)}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{U}_i, \boldsymbol{\mu}_{-\boldsymbol{U}_i})} \mathrm{d}\boldsymbol{\mu}_{-\boldsymbol{U}_i} \mathrm{d}\boldsymbol{U}_i \\ &\approx \int \frac{p(\boldsymbol{U}_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{\mu}_{-\boldsymbol{U}_i}) p(\boldsymbol{\mu}_{-\boldsymbol{U}_i} | \mathcal{D}, \boldsymbol{\theta}_w)}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{U}_i, \boldsymbol{\mu}_{-\boldsymbol{U}_i})} \mathrm{d}\boldsymbol{\mu}_{-\boldsymbol{U}_i} \mathrm{d}\boldsymbol{U}_i. \end{aligned}$$

We use the notation $\boldsymbol{\mu} = (\boldsymbol{U}_i, \boldsymbol{\mu}_{-\boldsymbol{U}_i})^\intercal$ to separate the random effects $\boldsymbol{U}_i$ that strictly depend on the loan i from the rest of the parameters $\boldsymbol{\mu}_{-\boldsymbol{U}_i}$.

Let $\boldsymbol{\mu}_{-\boldsymbol{U}_i}^{(r,w)}$ denotes the rth realisation of the approximated posterior sample with $r = 1, \ldots, R$, then $p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})^{-1}$ from Equation 5.9, can be estimated as

$$\begin{aligned} p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})^{-1} &\approx \sum_w \frac{\hat{p}(\boldsymbol{\theta}_w | \mathcal{D})}{\hat{p}(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i}, \boldsymbol{\theta}_w)} \Delta_w \\ &\approx \sum_w \hat{p}(\boldsymbol{\theta}_w | \mathcal{D}) \Delta_w \int \frac{p(\boldsymbol{U}_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{\mu}_{-\boldsymbol{U}_i}) p(\boldsymbol{\mu}_{-\boldsymbol{U}_i} | \mathcal{D}, \boldsymbol{\theta}_w)}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{U}_i, \boldsymbol{\mu}_{-\boldsymbol{U}_i})} \mathrm{d}\boldsymbol{\mu}_{-\boldsymbol{U}_i} \mathrm{d}\boldsymbol{U}_i \\ &\approx \sum_w \hat{p}(\boldsymbol{\theta}_w | \mathcal{D}) \Delta_w \left[\frac{1}{R} \sum_r \int \frac{p(\boldsymbol{U}_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{\mu}_{-\boldsymbol{U}_i}^{(r,w)})}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}_w, \boldsymbol{U}_i, \boldsymbol{\mu}_{-\boldsymbol{U}_i}^{(r,w)})} \mathrm{d}\boldsymbol{U}_i \right]. \end{aligned}$$

Furthermore, the integral can be calculated, for instance, with empirical Bayes or the Laplace method (Tierney and Kadane, 1986). Whichever method is used

to calculate the integral, denote this term as $h_i(\boldsymbol{\theta}_w, \boldsymbol{\mu}_{-\boldsymbol{U}_i}^{(r,w)}|t)$. Hence, cvDCL$(t)$ can be estimated as

$$\widehat{\text{cvDCL}}(t)^{INLA} = \frac{1}{N_t}\sum_{i=1}^{N} I(T_i > t)\times \\ \times \log\left\{\sum_w \hat{p}(\boldsymbol{\theta}_w|\mathcal{D})\Delta_w \left[\frac{1}{R}\sum_r h_i(\boldsymbol{\theta}_w, \boldsymbol{\mu}_{-\boldsymbol{U}_i}^{(r,w)}|t)\right]\right\}. \quad (5.10)$$

To get an estimate of the Monte Carlo variance of Equation 5.10, we use what is known as the *Delta method* (Ver Hoef, 2012). This method approximates a function of random variables using a Taylor series expansion around the means. In our case, we can identify the random variables as $h_{iwr|t} = h_i(\boldsymbol{\theta}_w, \boldsymbol{\mu}_{-\boldsymbol{U}_i}^{(r,w)}|t)$ which are independent for all the loans i, the integration points w and the realisations r. Denote $m_{iw|t} = \text{E}(h_{iwr|t})$ and $\sigma^2_{iw|t} = \text{Var}(h_{iwr|t})$ and their estimations, respectively, as $\hat{m}_{iw|t} = \frac{1}{R}\sum_r \hat{h}_{iwr|t}$ and $\hat{\sigma}^2_{iw|t} = \frac{1}{R-1}\sum_r(\hat{h}_{iwr|t} - \hat{m}_{iw|t})^2$. Then, the first order approximation of $\widehat{\text{cvDCL}}(t)^{INLA}$ as a function of the vector $\boldsymbol{h}_{|t} = \{h_{iwr|t}\}$ around the vector of means $\boldsymbol{m}_{|t} = \{m_{iw|t}\}$ is

$$\widehat{\text{cvDCL}}(t)^{INLA} = g(\boldsymbol{h}_{|t}) \approx g(\boldsymbol{m}_{|t}) + \sum_{i,w,r}(h_{iwr|t} - m_{iw|t})\left.\frac{\partial g}{\partial h_{iwr|t}}\right|_{\boldsymbol{h}_{|t}=\boldsymbol{m}_{|t}}. \quad (5.11)$$

Note that by construction $\text{E}(g(\boldsymbol{h}_{|t})) \approx g(\boldsymbol{m}_{|t})$. Moreover, the partial derivative terms follow

$$\left.\frac{\partial g}{\partial h_{iwr|t}}\right|_{\boldsymbol{h}_{|t}=\boldsymbol{m}_{|t}} = \frac{I(T_i > t)}{N_t}\frac{\hat{p}(\boldsymbol{\theta}_w|\mathcal{D})\Delta_w}{R\sum_w \hat{p}(\boldsymbol{\theta}_w|\mathcal{D})\Delta_w \hat{m}_{iw|t}}.$$

We compute the variance of the expression in Equation 5.11. Given that the terms $h_{iwr|t}$ are independent and using the partial derivative expression from above, it follows that

$$\begin{aligned}\text{Var}(\widehat{\text{cvDCL}}(t)^{INLA}) &\approx \sum_{i,w,r}\text{Var}(h_{iwr|t} - m_{iw|t})\left(\left.\frac{\partial g}{\partial h_{iwr|t}}\right|_{\boldsymbol{h}_{|t}=\boldsymbol{m}_{|t}}\right)^2 \\ &= R\sum_{i,w}\hat{\sigma}^2_{iw|t}\left(\left.\frac{\partial g}{\partial h_{iwr|t}}\right|_{\boldsymbol{h}_{|t}=\boldsymbol{m}_{|t}}\right)^2 \\ &= \frac{1}{N_t^2 R}\sum_i I(T_i > t)\frac{\sum_w \hat{\sigma}^2_{iw|t}(\hat{p}(\boldsymbol{\theta}_w|\mathcal{D})\Delta_w)^2}{(\sum_w \hat{p}(\boldsymbol{\theta}_w|\mathcal{D})\Delta_w \hat{m}_{iw|t})^2}.\end{aligned} \quad (5.12)$$

The estimation of cvDCL(t) follows Equation 5.10 and its variance Equation 5.12.

Originally, Rizopoulos et al. (2016) propose to estimate the cvDCL(t) with posterior samples from an MCMC simulation as detailed in Appendix C.1. In order to compare how suitable our estimate is relative to the author's, we perform a comparison analysis for simulated datasets described in Appendix C.2.

5.3 Full prepayment prediction on US mortgages

5.3.1 Data

We use the Single Family Loan-Level Dataset publicly provided by Freddie Mac[3]. This mortgage dataset has loan-level granularity with application covariates and monthly performance information and is periodically updated. The training dataset includes the loans granted from June, 2015 to November, 2015 and followed until December 2019, hence the maximum period of performance records per loan is 4.5 years (54 months). This corresponds to 57,258 borrowers with a total of 2,559,056 observations. Out of these borrowers, 16,239 of them full prepaid their mortgage loans during the study period. Figure 5.1 shows the distribution of the full prepayment events in time.

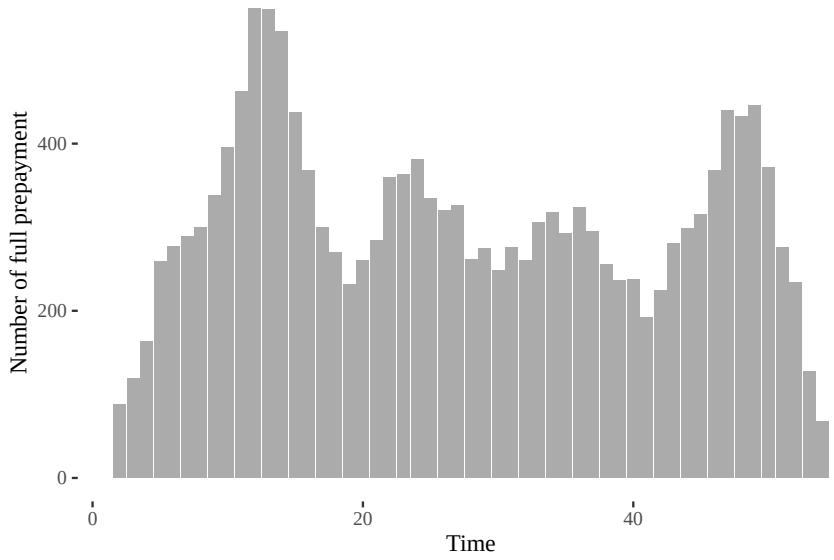

Figure 5.1: Distribution of the full prepayment events in time.

The time-fixed covariates included in the survival process and described in Equa-

[3]Visit `https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset`.

tion 5.2 by the vector $\boldsymbol{z}_i$ are the following[4]

- **cltv** is the loan-to-value ratio based on the original mortgage loan amount plus any other mortgage loan amount divided by the property's purchase price.
- **orig_upb** is the original unpaid principal balance of the mortgage on the note date.
- **cnt_units** denotes whether the mortgage is a one- (= 1, 93% of the loans), or more than one-unit property (= 0, 7% of the loans).
- **dti** is the debt to income ratio. It corresponds to the borrower's monthly debt payments divided by the total monthly income used to underwrite the loan.
- **int_rt** is the interest rate given at the origination of the credit.
- **term** corresponds to the number of scheduled monthly mortgage payments. It is divided between short-term loans, with terms less than or equal to 15 years (= 0, 19% of the loans) and long-term loans, with terms greater than 15 years (= 15, 81% the loans).
- **loan_purpose** indicates whether the mortgage loan purpose is a cash-out refinance loan (= 0, 22% of the loans)[5], no cash-out refinance loan (= N, 25% of the loans) or purchase (= P, 53% of the loans).
- **cnt_borr** is the number of borrowers obligated to repay the mortgage. Either one borrower (= 0, 48% of the loans) or more than one (= 2, 52% of the loans).

Table 5.1 shows descriptive statistics of the numeric covariates defined above. As a pre-processing step, these variables are standardised to have a zero-mean and standard deviation of 1.

Concerning the longitudinal outcome, we are interested in a variable that can be simple and indicative of early repayments. For example, in Chapter 4, we saw for a consumer loans dataset that those borrowers who have paid more than the amount that was due are correlated with the prepayment event. For this mortgage loan dataset, we follow the same rationale of looking for a candidate variable that measures the distance between what was paid and what was due. Moreover, in

[4]See, for example, Wang et al. (2020); Hu and Zhou (2019) who use this dataset in similar contexts.

[5]A cash-out refinance mortgage loan is a loan in which the use of the amount is not limited to specific purposes.

Covariate	N	Mean	SD	$Q_{2.5\%}$	$Q_{25\%}$	$Q_{50\%}$	$Q_{75\%}$	$Q_{95\%}$
cltv	57258	73.50	16.97	38.00	65.00	79.00	85.00	95.00
orig_upb*	57258	256.32	121.87	88.00	161.00	241.00	336.00	475.00
dti	57258	34.87	9.14	19.00	28.00	36.00	42.00	48.00
int_rt	57258	3.93	0.44	3.00	3.75	4.00	4.25	4.62

***1,000 USD.**

Table 5.1: Descriptive statistics for numeric covariates in the dataset.

terms of simplicity, we are looking for a variable that can be described by a simple functional structure, such as a linear relationship. This simplifies the longitudinal structure and, therefore, the model estimation.

The following variable encompasses those above. Assume that for a generic loan, we denote the interest rate given at origination as i with a monthly instalment equal to A. Then, the sum of the total amount paid until time t, including the capitalisation of the inflows, is $A + A(1+i) + \ldots + A(1+i)^{t-1}$. Since the interest rate is commonly low, especially for mortgage loans (the 95% quantile for our dataset is 4.62%, which is equivalent to a monthly interest rate of 0.0038%. See Table 5.1), we can do a first-order Taylor series expansion of the previous expression with respect to i around zero, which is simple $At + At(t-1)i/2$. Note that for the first periods, the linear term of this expression dominates, which is what we are looking for.

To make the longitudinal outcome more comparable among different loans, we finally define the following variable $y_t = \sum_{s=0}^{t-1}(1+i)^s/T$, where T is the length of the study (54 months in our case), and it only has a scaling purpose. Note that the following expression also holds for $y_t = \frac{(1+i)^t-1}{iT}$. If the total amount paid by the borrower is greater than what it was due at time t, then the observed y_t should have a larger slope than the theoretical curve.

We use basic instalment relationships to calculate the observed y_t concerning the unpaid principal balance. Denote the original unpaid principal balance as P_0, the current unpaid principal balance at time t as P_t and the term of the loan as M, then an equivalent expression for y_t is

$$y_t = \frac{(P_0 - P_t)}{P_0} \frac{(1+i)^M - 1}{iT}.$$

Figure 5.2 shows the aforementioned longitudinal outcome for our dataset. To

facilitate visualisation, we have highlighted in red dashed line loans that evidenced the event of full prepayment, and in blue dotted line, loans that did not.

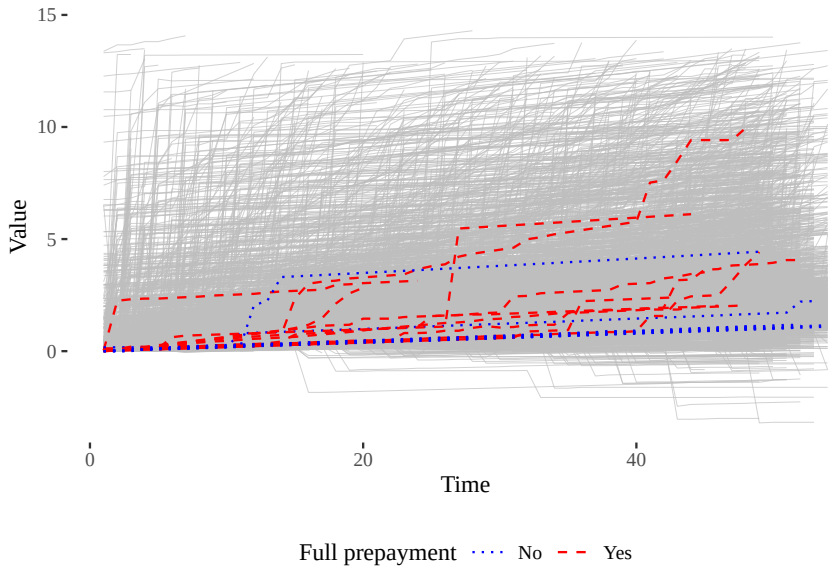

Figure 5.2: Evolution of the longitudinal outcomes. For visual purposes, we highlight borrowers who full prepaid the loan (dashed line in red) and borrowers that are censored (dotted line in blue).

Regarding geographical information, properties are located in 8 states: New York, New Jersey, Connecticut, Massachusetts, Rhode Island, Maine, New Hampshire and Vermont. These states are divided into a total of 123 areas given by the first three digits of the postcode. The number of loans distributed among these areas is shown in the map depicted in Figure 5.3. In addition, Figure 5.4 indicates the corresponding full prepayment rates, calculated as the total number of events divided by the number of granted loans in each area. From this last figure, although the rates include all events regardless of when they occurred, spatial clustering is observed and can be considered a first check to support the inclusion of spatial effects.

5.3.2 Models and results

Following the methodology described in Section 5.2.2, we estimate three specifications of joint models. All three include the same time-fixed covariates described in Section 5.3.1 and also the same structure of the longitudinal outcome (see below). Rather the differences come from the assumed baseline terms $\nu_{a,s}$ (see

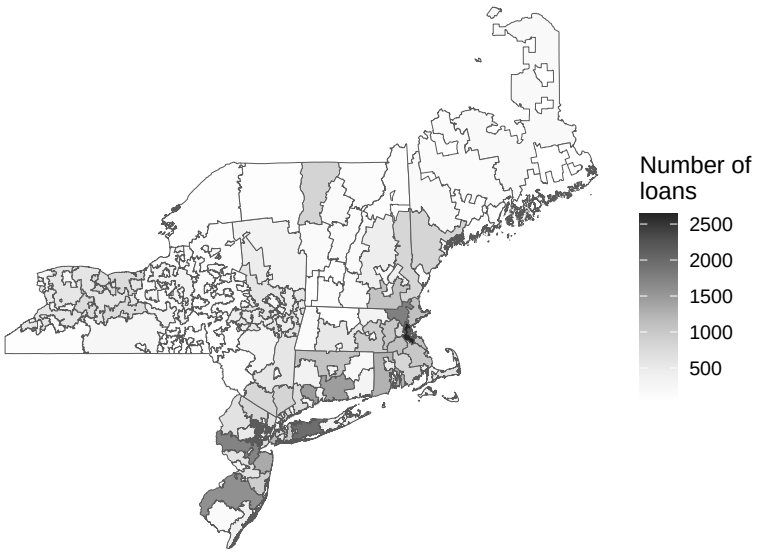

Figure 5.3: Number of loans distributed by area.

Equation 5.2). Concretely, the longitudinal outcome follows

$$\begin{aligned}(Y_{i,s}|\eta_{Yi,s},\tau_Y) &\sim N(\eta_{Yi,s},\tau_Y^{-1})\\ \eta_{Yi,s} &= \beta_{01}+\beta_{11}s+U_{0i}+U_{1i}s\\ (U_{0i},U_{1i})^\intercal &\sim N_2(0,Q_{\boldsymbol{U}}^{-1}),\end{aligned} \tag{5.13}$$

where the covariance matrix $Q_{\boldsymbol{U}}^{-1}$ is parameterised via marginal precisions τ_{U_0} and τ_{U_1}, and the pairwise correlation ρ_{01} as follows

$$Q_{\boldsymbol{U}}^{-1} = \begin{pmatrix} 1/\tau_{U_0} & \rho_{01}/\sqrt{\tau_{U_0}\tau_{U_1}} \\ \rho_{01}/\sqrt{\tau_{U_0}\tau_{U_1}} & 1/\tau_{U_1}\end{pmatrix}. \tag{5.14}$$

The mixed-effect model from Equation 5.13 is known as intercept-slope random effects. This specification is justified by the fact that the longitudinal outcome approximates a linear trend when the interest rate is low, as shown in Section 5.3.1. Moreover, the survival process for the three models follow

$$\begin{aligned}(X_{i,s}|X_{i,s-1}=0,\eta_{Xi,s}) &\sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{Xi,s}))\\ \eta_{Xi,s} &= \nu_{a_i,s}+\boldsymbol{z}_i^\intercal\boldsymbol{\beta}_2+\lambda(U_{0i}+U_{1i}s),\end{aligned} \tag{5.15}$$

where $\nu_{a_i,s}$ is the baseline risk. Table 5.2 describes the three models' specification for $\nu_{a,s}$. Note that M_1 is a joint model in discrete time, similar to the ones estimated in Chapter 3 (univariate, without autoregressive terms). M_2, however, is new to the literature in the sense that, as far as we know, there is no study of a

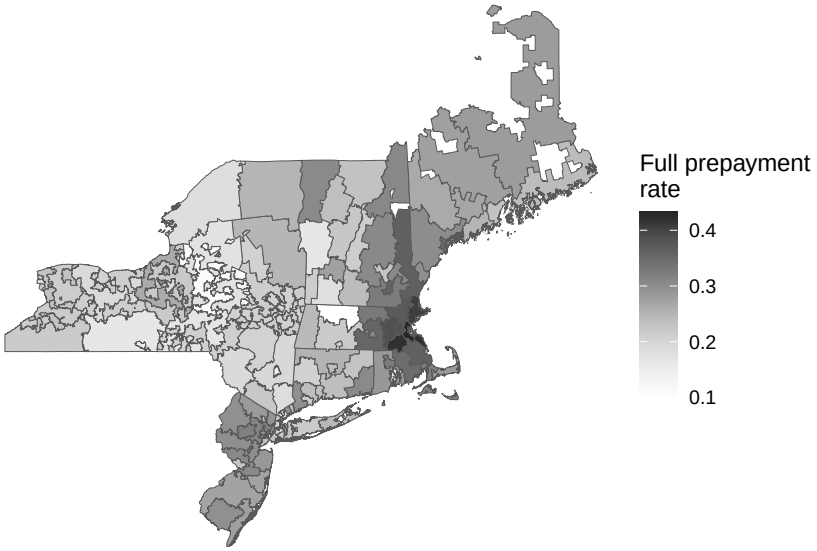

Figure 5.4: Full prepayment rate distributed by area.

joint model with spatial main effects in discrete time. The work of Martins et al. (2016) can be seen as the closest one, which presents a joint model that includes the spatial main effects in a Weibull survival sub-model. Finally, M_3 is the model that encompasses all the effects, that is, the temporal and spatial main effects as well as the interactions.

Id	Temporal Effects	Spatial Effects	S-T Interactions	$\nu_{a,s}$
M_1	Yes	No	No	$\nu_0 + v_s$
M_2	Yes	Yes	No	$\nu_0 + v_s + u_a$
M_3	Yes	Yes	Yes	$\nu_0 + v_s + u_a + \delta_{a,s}$

Table 5.2: Specification of the joint models. M_1 only includes the temporal effects in the baseline hazard. M_2 has both the temporal and spatial main effects, and M_3 includes the interactions among them apart from both main effects.

Table 5.3 shows the parameter estimates for the three models. We observe that the parameters strictly associated with the longitudinal outcome, β_{01} and β_{11}, are consistent among M_1, M_2 and M_3. However, we notice differences in the covariates associated with the survival process. For instance, the coefficient related to *cltv* under the estimation of M_1 is 0.301 and its 95% posterior credible interval does not include zero. The positive sign suggests that the higher the *cltv*, the greater the probability of prepaying in full. Yet, when estimated under specifications M_2 and M_3, although the sign remains positive, the effect of this

covariate decreases and is not as significant as before.

In the same line, we notice that the effect of *dti* for M_1 shows a negative relationship with the prepayment. Similar results were found in Chapter 4 for the consumer loans dataset. However, when we include the spatial effects, either with M_2 or M_3, the relation of high *dti* with a low probability of full prepayment is not entirely conclusive, even shifting the posterior marginals to the positive values.

For the other covariates, we found agreeing results among the three models. For example, the original unpaid principal balance, *orig_upb*, shows a positive relationship with the prepayment, which is also supported by the prepayment models from Chapter 4. Moreover, it is less likely to prepay in full if the mortgage is more than one-unit property (*cnt_units*), if its term is longer than 15 years (*term_g15*), if the number of borrowers is greater than 1 (*cnt_borr2*) or if the purpose of the loan is to purchase rather than refinance (*loan_purpose*).

	M_1			M_2			M_3		
	Mean	2.5%	97.25%	Mean	2.5%	97.25%	Mean	2.5%	97.25%
β_{01}	0.011	0.008	0.013	0.011	0.008	0.013	0.011	0.008	0.013
β_{11}	0.027	0.026	0.027	0.027	0.026	0.027	0.027	0.026	0.027
ν_0	-8.247	-8.451	-8.043	-8.590	-8.798	-8.382	-8.541	-8.748	-8.334
cltv	0.301	0.081	0.520	0.114	-0.120	0.347	0.117	-0.115	0.349
orig_upb	0.143	0.128	0.159	0.153	0.134	0.172	0.153	0.134	0.172
cnt_units1	0.439	0.367	0.510	0.425	0.351	0.499	0.408	0.334	0.482
dti	-0.109	-0.222	0.005	0.024	-0.091	0.139	0.018	-0.097	0.132
int_rt	0.794	0.743	0.845	0.869	0.817	0.921	0.846	0.794	0.898
term_g15	-0.458	-0.518	-0.399	-0.531	-0.591	-0.470	-0.518	-0.579	-0.458
loan_purposeN	0.028	-0.018	0.074	-0.005	-0.051	0.041	-0.010	-0.056	0.036
loan_purposeP	-0.137	-0.180	-0.093	-0.100	-0.144	-0.057	-0.109	-0.153	-0.066
cnt_borr2	-0.060	-0.091	-0.028	-0.061	-0.093	-0.029	-0.063	-0.095	-0.031
λ	0.201	0.188	0.211	0.146	0.139	0.156	0.171	0.162	0.184
τ_Y	34.713	34.653	34.777	34.715	34.656	34.783	34.784	34.730	34.826
τ_{U_0}	9.651	9.568	9.752	9.627	9.557	9.716	9.870	9.767	9.956
τ_{U_1}	895.736	885.777	904.230	887.034	874.223	898.767	880.029	871.477	890.400
ρ_{01}	-0.067	-0.076	-0.058	-0.052	-0.060	-0.046	-0.056	-0.064	-0.051
τ_v	3.004	2.245	3.653	2.435	1.369	3.375	0.802	0.518	1.094
τ_u				16.667	12.163	25.503	14.777	12.806	17.464
τ_δ							104.613	57.088	196.704

Table 5.3: Parameter estimations of models M_1, M_2 and M_3.

The interest rate granted at origination, *int_rt*, is expected to play an essential role in the decision of full prepayment since if the reference interest rates fall compared to the one granted, it is more attractive to renegotiate the credit. As seen in Table 5.3 for all three models, its effect is positive, which suggests that the higher the interest rate given at origination, the greater the probability of full prepayment. This has also been seen in Chapter 4. However, when we include the spatial effects, we note that the associated coefficient also increases.

Regarding λ, the parameter that associates the random effects of the longitudinal outcome and the survival process, we observe that the three models estimate a significant positive effect, as expected since the more is paid off from what is owed, the more likely it is to prepay in full. However, the magnitude of the estimate differs among the models. The largest one is due to M_1 with a mean of 0.201. When we add the spatial main effects in M_2, we see a decrease of the mean to 0.146. Yet, when we add the spatio-temporal interactions in M_3, we observe a value in between, with a mean of 0.171.

Furthermore, we obtained similar results among the three models regarding the hyperparameters associated with the longitudinal outcome. Namely, the precision of the innovations τ_Y and the elements τ_{U_0}, τ_{U_1} and ρ_{01} of the precision matrix $Q_{\boldsymbol{U}}$ (see parametrisation in Equation 5.14). However, the precision of the temporal main effects τ_v changes among the three models. We see a mean of 3.004, 2.435 and 0.802 for models M_1, M_2 and M_3, respectively. That raises the question of how different the estimated temporal main effects for each model are. Figure 5.5 shows the estimated temporal main effects for the three models. Models M_1 and M_2 overlap for much of the study period, and M_3 shows some differences, in particular for the first periods, but, overall, the effect of the three models is fairly comparable.

To compare the performance of the models, we follow the procedure described in Section 5.2.3. We estimate the $\widehat{\text{cvDCL}}(t)^{INLA}$ (Equation 5.10) for six evaluation times t, ranging from 12 to 42 months with an increment of 6 months. The results are shown in Table 5.4 (we deliberately omit the word "INLA" to shorten the notation). N_t is the number of borrowers at risk, and the values in brackets are the estimates of Monte Carlo standard deviation derived from Equation 5.12. It is worth noting that the metric value should be compared across the models for one value of t, that is, all the values that belong to the same row since, between

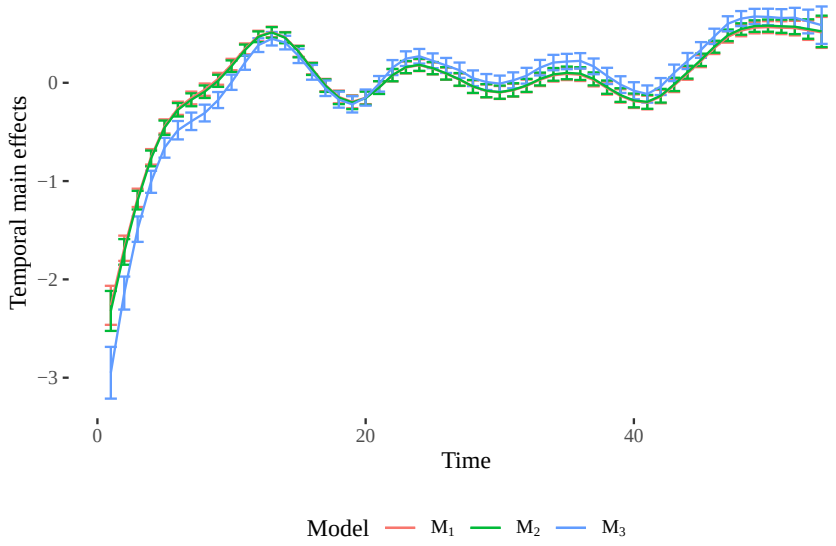

Figure 5.5: Temporal main effects estimated by the three models. The error bars represent the estimated 95% credible intervals.

the rows, there is an evident overlap of datasets. From the table we observe that both M_2 and M_3 outperform M_1. Adding the latent spatial component can increase the model's performance for this dataset. However, when we compare models M_2 and M_3, that is when we add on top of the spatial main effects, the spatio-temporal interactions, the improvements are not as clear as before.

	N_t	M_1	M_2	M_3
$\widehat{\text{cvDCL}}(t=12)$	53963	1.4438 (5.69e-06)	1.4244 (1.03e-05)	1.4304 (5.72e-05)
$\widehat{\text{cvDCL}}(t=18)$	51489	1.2231 (4.01e-06)	1.2143 (7.57e-06)	1.2165 (2.20e-05)
$\widehat{\text{cvDCL}}(t=24)$	49607	1.0349 (3.58e-06)	1.0306 (6.91e-06)	1.0310 (1.35e-05)
$\widehat{\text{cvDCL}}(t=30)$	47839	0.8472 (3.27e-06)	0.8450 (6.40e-06)	0.8448 (1.15e-05)
$\widehat{\text{cvDCL}}(t=36)$	46059	0.6453 (2.91e-06)	0.6438 (5.68e-06)	0.6439 (1.08e-05)
$\widehat{\text{cvDCL}}(t=42)$	44611	0.4656 (2.52e-06)	0.4644 (4.96e-06)	0.4648 (1.00e-05)

Table 5.4: Comparison of model performance. The value in brackets is an estimate of the Monte Carlo standard deviation.

To further explore the analysis, we assign the overall $\widehat{\text{cvDCL}}(t)$ to the corresponding area and compare from which areas the major gains are obtained for models M_2 and M_3 with respect to model M_1. Figure 5.6 shows two maps for the segmented $\widehat{\text{cvDCL}}$ evaluated at $t=12$. The left one corresponds to the difference between M_2 and M_1 $(M_2 - M_1)$, and the right one to $M_3 - M_1$. From both maps, we observe that the major contributions to the overall metric mainly come from

the middle-left (west, Rochester area) and middle-right parts (east, Boston area) of the maps. These differences are increased for model M_2.

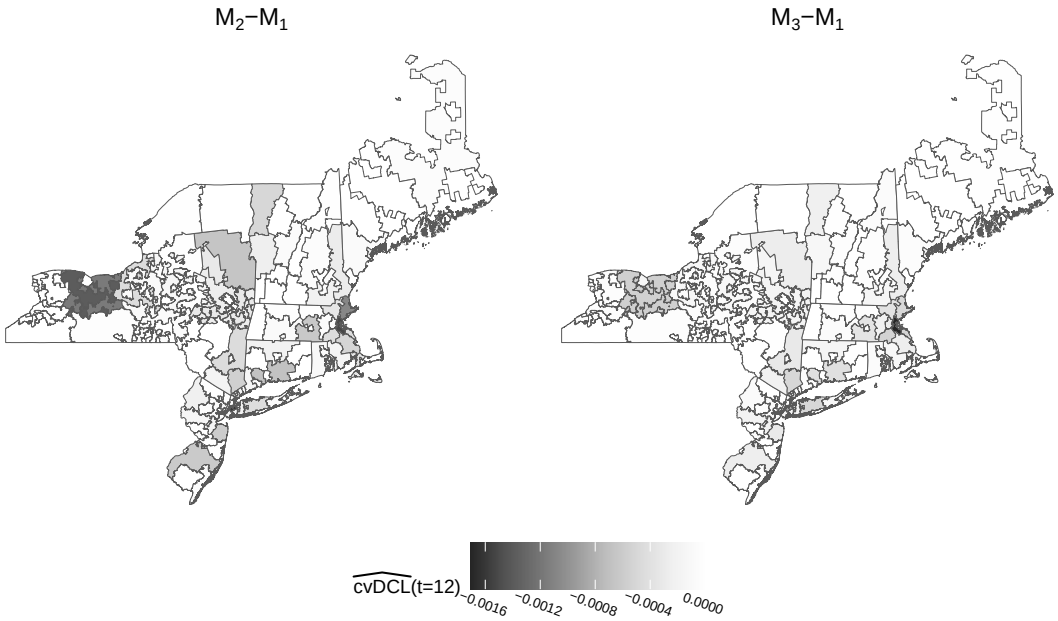

Figure 5.6: Difference between the $\widehat{\text{cvDCL}}(t = 12)$ for models M_2 and M_3 with respect to M_1 and segmented by area.

Moreover, when we choose a different evaluation time, for instance, $t = 24$ (see Figure 5.7), now the contributions coming from the Rochester area are not as important as for $t = 12$. Rather the differences come from areas of New Jersey, New York City and Boston. Therefore, when we include spatial effects, we see that the consistent improvements in the performance evaluated in different periods are not exclusively attributed to a particular area.

5.4 Discussion

Chapters 3 and 4 show that the joint model approach has advantages over the survival approaches commonly used for credit-related applications (Stepanova and Thomas, 2002; Bellotti and Crook, 2014; Wang et al., 2020). In particular, we have explored this approach's flexibility to model one or more longitudinal outcomes with different specifications. In this chapter, we keep exploring the joint model flexibility, but now we focus on finding better representations of

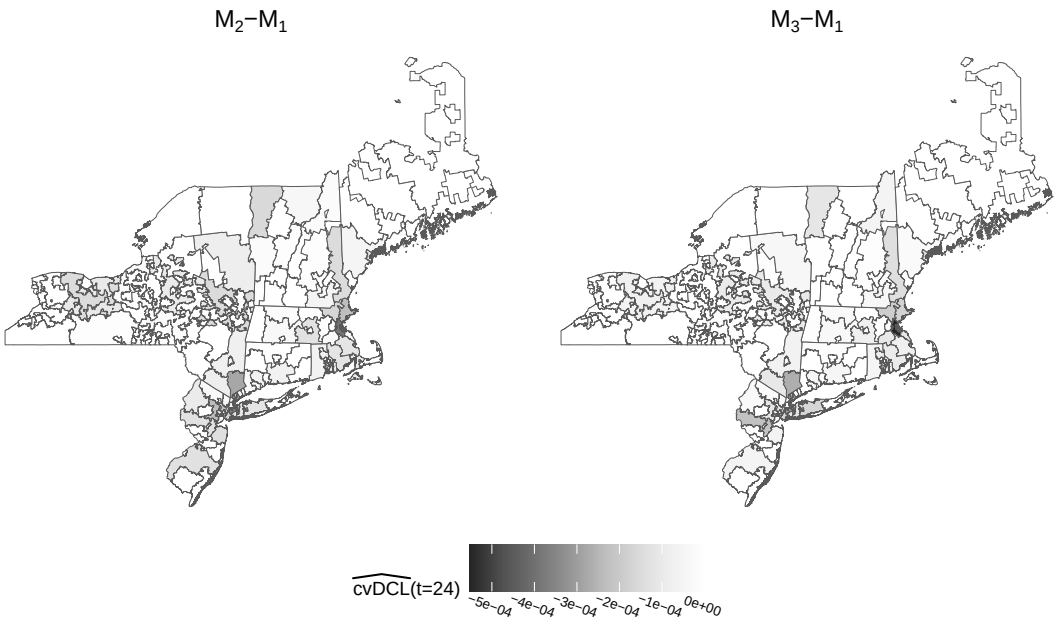

Figure 5.7: Difference between the $\widehat{\text{cvDCL}}(t = 24)$ for models M_2 and M_3 with respect to M_1 and segmented by area.

the survival predictor rather than in the longitudinal part. Concretely, we are interested in including spatial and spatio-temporal effects in the baseline hazard and studying how this can change prediction performance for a prepayment model for US mortgages. This decision is supported, first, by the evidence that including spatial effects in survival models can lead to better predictions (Calabrese and Crook, 2020; Medina-Olivares et al., 2022b). And second, as we have mentioned before, by jointly modelling the longitudinal and survival processes, we have an appealing prediction framework for credit modelling.

In this chapter, we make four contributions to the literature. First, we present the Spatio-Temporal Joint Model (STJM), a joint model formulated in discrete time that includes a flexible baseline hazard in the survival predictor. This baseline hazard is decomposed between temporal and spatial main effects and the interactions among them. For this latter, we leverage the approach from Clayton (1996) in which the structure matrix is built by the Kronecker product of the main effects structure matrices. Moreover, we follow the Goicoa et al. (2018) approach to get appropriate identifiability constraints by using spectral decomposition over the structure matrices.

Second, to estimate the STJM in a large dataset, we formulate the model using the INLA methodology (Rue et al., 2009) and implement it in the `R-INLA` package (`https://www.r-inla.org/`). This implementation allows us to scale the model to a dataset with 57,258 borrowers with 2,559,056 total observations. As far as we know, this is the largest sample size used in a joint model application.

Third, we introduce a modified version of the *cross-validated Dynamic Conditional Likelihood* proposed by Rizopoulos et al. (2016). Our version takes advantage of the estimations already performed by the INLA methodology, which is not based on posterior MCMC samples as the original version, reducing computational costs. We compare the original and the proposed versions by a simulation study that demonstrates adequate results (see Appendix C.2).

Fourth, we apply the proposed approach to predict the full prepayment event in US mortgage loans. The analysis consists of three models that measure (1) only the temporal main effect (M_1), (2) the addition of the temporal and spatial main effects (M_2) and (3) the composition of both main effects plus their interactions (M_3). The parameter estimates generally agree between the three models. However, a notable difference comes from the covariate debt to income ratio *dti* (sum of the borrower's monthly debt payments divided by the monthly income). When no spatial effects are included, the parameter estimate associates greater values with a lower probability of prepayment, but this relation no longer holds when we have the spatial component.

We also find that spatial effects can consistently improve the joint model's prediction performance for different evaluation times. Furthermore, we see that these improvements are not strictly associated with a particular area when we compare the performance evaluated at other times. However, when the spatio-temporal interactions are included, the performance gains are less clear to the model that does not have them.

This study yields exciting results and undoubtedly opens the way for further research. For instance, to study the implications in both model estimation and predictions when we include external TVCs in the survival process. Examples of these TVCs are the macroeconomic variables, and the idea is to explore how the changes in the general conditions of the economy affect the model performance. This has been done previously in credit survival models (Bellotti and Crook,

2009a; Djeundje and Crook, 2018; Dirick et al., 2019). Since these TVCs are external, we can assume that the occurrence of a particular event does not influence their paths; therefore, no borrower-specific longitudinal model is needed for these covariates. That can lead to a more general joint model framework that makes the best of individual predictions, including the economic conjuncture.

Chapter 6

Concluding Remarks

6.1 Summary

Survival models are appealing for modelling credit events. Unlike classification approaches, they can provide information about when the event is likely to occur. Moreover, they present a flexible framework to include time-varying covariates (TVCs) and censored observations. Similarly, the inclusion of TVCs has been shown to improve prediction performance. However, the survival approaches commonly used in this context cannot handle potential endogeneity in the TVCs concerning the event. Also, they do not offer a prediction framework incorporating the synchronised mutual evolution between survival processes and TVCs. As a result, the standard practices for building survival models for predictive purposes are either to lag the TVCs with respect to the event's occurrence or to carry their last observations forward. Unfortunately, neither appears to be optimal for prediction.

This thesis's general objective addresses whether handling for potential endogeneity on the TVCs can improve the predictive performance in the survival context. To this mean, we explore a new way to conduct survival analysis, assuming that survival time and the endogenous TVCs (also termed longitudinal outcome) are stochastic processes linked through a latent structure. Initially developed in medical research, this novel approach is known as the joint model of longitudinal and survival data. Although we can readily find analogies between medical and credit risk applications, they have significant distinctions. Therefore, we propose a series

of innovations to make the approach suitable for credit-related applications.

Specifically, these distinctions relate to the nature of survival time, the evolution of the endogenous TVCs, scalability of the joint models to large samples and the type of information available in credit data.

In terms of the nature of survival time, credit loan data are usually recorded every month, and events, such as the default or the full prepayment, are defined based on the realised payment on the billing date (e.g. end of the month), making the time to event intrinsically discrete. That is not the case in medical applications, where most literature assumes time as continuous. Moreover, we commonly have performance variables in credit data, such as unpaid principal balance. These records are expected to be highly correlated to the previous ones since we have scheduled curves, and most borrowers comply with them. Therefore, any estimated deviation from it can be used as an early warning of the probability of the event occurring.

Regarding scalability, the standard estimation procedures of joint models are computationally expensive. Hence, to apply these models in environments where datasets are large and sometimes with more than one longitudinal outcome, we need alternative ways both to estimate them and to make individual survival predictions. Finally, in credit datasets, access to variables such as the spatial location of the borrower is not uncommon, and we can then leverage this information into the modelling framework.

The contributions are developed in Chapter 3, 4 and 5.

In Chapter 3, we make two main contributions, one from an applied perspective and the other methodological. In terms of application, we introduce joint modelling for discrete survival data in the credit context. Notably, we take advantage of the flexibility of the approach and build five joint models to predict default in US mortgages. The models have different specifications for the longitudinal outcomes and the link function between both processes. From a methodological point of view, we propose to include autoregressive terms in the longitudinal submodel. This innovation is motivated by the autocorrelation found in the data and the predictive implications that these additional terms can have. We study the proposed joint model via simulation, recovering the true parameter values for different settings. The empirical results show that we can improve the model's

discrimination performance by including autoregressive terms.

In Chapter 4, we present four contributions: two methodological and two empirical. From the methodological point of view, first, we propose a computationally efficient implementation of a joint model with two longitudinal outcomes. The model is framed using the INLA methodology, which allows us to scale it to credit data environments. To illustrate the proposed implementation, we conduct a simulation analysis showing satisfactory recovery of the true parameter values considering different sample sizes. Second, we offer a methodology for individual survival predictions based on the Laplace method. The proposal is theoretically more accurate than empirical Bayes techniques and, unlike simulation-based approaches, is meant to be applied in situations where several out-of-sample individuals need to be evaluated.

From an empirical standpoint, we apply the multivariate joint modelling approach in credit literature for the first time. Specifically, we build predictive joint models for full prepayment events aimed at consumer loans in Germany. In addition, we show that the multivariate version can outperform traditional survival approaches in calibration and discrimination metrics in out-of-sample and out-of-time settings.

In Chapter 5 we delve into finding better representation in the survival predictor leveraging the spatial information at hand. To this extent, we make four contributions. First, we introduce the Spatio-Temporal Joint Model (STJM), which has a flexible baseline hazard decomposed into temporal and spatial main effects and the interaction among them. Following Clayton (1996), we represent the interaction as the Kronecker product of the main effects structure matrices. To make the STJM well-posed, we set identifiability constraints derived from spectral decomposition, as suggested by Goicoa et al. (2018). This approach allows us to study the survival effect due to the spatio-temporal correlation between events happening within nearby areas.

The second contribution relates to the scalability of the STJM implementation. As in Chapter 4, we formulate the approach using INLA, which permits us to apply it in a training sample of 57,258 borrowers, representing 2,559,056 observations, the largest in this context as far as we know.

Third, we introduce a new implementation of the *cross-validated Dynamic Con-*

ditional Likelihood (cvDCL) proposed by Rizopoulos et al. (2016). The cvDCL is a cross-entropy estimate of the survival outcome's cross-validatory posterior predictive conditional density and lets us compare the predictive performance among different models. The advantage of our proposal is computational since we do not rely on simulation-based schemes as originally introduced. Rather we use the estimates already computed by INLA.

And finally, from an empirical perspective, we apply the STJM to predict full prepayment in US mortgage loans, obtaining that the inclusion of spatial effects consistently improves the performance for different evaluation periods. In contrast, the spatio-temporal interactions do not provide significant gains when considering the main effects.

6.2 Limitations

Three limitations associated with the joint modelling approaches used throughout this thesis are important to note.

First, we have seen that the estimation is computationally challenging compared to standard survival approaches. The main reason, in the discrete survival case, is the presence of random effects to represent the association between survival and longitudinal processes[1]. Maximum likelihood estimations comprise the computation of intractable integrals when marginalising over these terms (Equation 2.11), requiring numerical integration techniques. On top, we need additional assumptions and computations to estimate standard deviations. On the other hand, the Bayesian paradigm offers a flexible framework without the need for asymptotic assumptions or highly customised estimation procedures. This path allowed us to include autoregressive terms and different association structures, as seen in Chapter 3. However, the inference is performed on the full posterior distribution, increasing the parameter space (Θ and $\boldsymbol{U}$). MCMC methods, such as NUTS, are appropriate in these cases but are also computationally demanding, especially in credit data environments.

Second, we rely on the INLA methodology to scale the joint modelling. As seen in Section 2.2.5, INLA defines a large Gaussian latent field $\boldsymbol{\mu}$ conditional on a

[1] In the continuous case, the survival function is an integral, rather than a simple multiplication, which adds additional numerical complexities (see Section 2.1.2).

set of hyperparameters $\boldsymbol{\theta}$. For the INLA methodology to work well, the number of hyperparameters should not be too large. As described, INLA builds an integration grid in the hyperparameter space; therefore, a larger-dimensional space not only increases the computational burden but also makes approximating the marginal posterior distributions based on a discrete grid more daunting. A direct consequence of this limitation, for example, would be to implement joint models including a more significant number of correlated random effects specific to the borrower since we increase the dimension of the precision matrix ($Q_{\boldsymbol{U}}$). INLA would currently allow us to consider normally distributed random effects with a maximum dimension of 5. However, the results of Chapter 4 show that considering the correlation between random effects of different longitudinal outcomes did not necessarily imply significant improvements. To this extent, one option is if we wanted to include more TVCs in the specification of the model, each with their corresponding borrower-specific random effects, we could constrain the hyperparameters of those corresponding correlations to be zero and hence reduce the dimensionality.

And finally, in Section 2.1.2, we describe three types of censored observations related to when the true event is known to have happened (left, interval or right-censoring). That is an important distinction when specifying the proper survival method. In our case, we have considered right-censoring, which is sensible for the data at hand. However, another relevant consideration in survival contexts relates to the censoring mechanism. We have assumed that censoring is non-informative to the survival time, i.e. the reasons that some borrowers were censored at time t are unrelated to the event. Informative censoring, on the other hand, happens when borrowers withdraw from the study due to reasons related to the event. An example of a situation where censoring would be informative in our context is when the lender sells a portfolio to a third party that comprises several loans close to default. In this case, the lender would selectively reduce the events, and the validity of the statistical analysis obtained through a joint model would be in check. Methods to deal with informative censoring in a joint model framework have been proposed (Papageorgiou and Rizopoulos, 2021). Sensitivity analysis, for example, emulates different scenarios for the association between censored and event observations so that judgments can be drawn that are not solely based on criteria subject to the observed data.

6.3 Recommendations for future research

As a new methodological approach in the credit context, we envision several lines of future studies.

In Chapter 3, we include autoregressive terms of order 1 in the application, obtaining improvements in discrimination compared to specifications that do not have these terms. A natural route would incorporate higher-order terms or even evaluate other specifications to address the same goal of handling serial correlation and its implications in the predictions. Examples of alternative specifications can be random walks of order 1 or 2, similar to those used to identify the baseline hazard function in Chapters 4 and 5. As in the version of the joint model with autoregressive terms, the temporal effects of each TVC associated with the borrower would be thought of as a replication of a stochastic process with shared hyperparameters, the same for the entire population and inferred from the data. With their appropriate constraints, both random walk specifications are relatively straightforward to implement with INLA. On the contrary, if we wanted to assume that the TVCs of each borrower incorporate a process that does not share the hyperparameters, INLA would be out of the choices due to the limitations we mentioned above.

Moreover, in all three chapters, we assume that the effect between the longitudinal and survival processes is represented by an association parameter λ and a function of the longitudinal predictor (see Equation 2.9). However, there may be situations where relaxing some of these assumptions or including terms that account for additional effects make sense. For instance, when the strength of the association between the TVCs and survival processes is considered time-dependent, i.e. $\lambda(t)$. Similar studies unrelated to joint models have been applied in the credit risk literature (e.g. Djeundje and Crook, 2019a; Calabrese and Crook, 2020; Leow and Crook, 2016). Some of the TVCs employed by the authors are potentially endogenous but have not been treated as such. Therefore, it would be interesting to investigate the value that joint models can bring in these circumstances. In the joint model literature, this specification has been introduced for continuous-time showing that the dynamic prediction can be improved (Andrinopoulou et al., 2018). The authors estimate the model via MCMC. A genuine question will be how well this approach scales to our needs or if we require alternative estimation

procedures. INLA would not handle this sort of specification. However, a recent promising method, *inlabru* (Bachl et al., 2019), extends the possibilities of INLA to more general nonlinear predictors, such as in the case of the joint model with a time-varying association parameter. The *inlabru* method adds a linearisation step in the INLA estimation pipeline and is available through the R package `inlabru`[2].

The *inlabru* method would also allow us to include a more flexible link function, such as asymmetric ones. When there is a rare event, such as the default in mortgage loans (Chapter 3), it has been shown that we can underestimate the probability of the event by using symmetric link functions, such as the logit (Calabrese and Osmetti, 2013). Based on the extreme value theory, some authors have proposed using the generalized extreme value (GEV) cumulative distribution function (Calabrese et al., 2016) to model the tail of the response curve for values close to 1 (rare event). The GEV distribution is flexible because it has a parameter that controls the shape of the tail. In the context of survival analysis for credit defaults, the GEV link function has shown better results than symmetric links (Calabrese and Crook, 2020). The joint model with a GEV link function could be, in principle, estimated through *inlabru*, where the parameter associated with the GEV distribution can be estimated thanks to the linearisation step.

Another promising avenue of research is when we want to model multiple types of events. For example, if we are interested in estimating the profit of a loan, we can consider default and prepayment events as competing risks that affect the final yield (Banasik et al., 1999; Bhattacharya et al., 2019). The traditional approach is cause-specific and proposes separate hazard regressions, one for each cause and considering the events associated with the other cause as censored observations. However, we can use a joint model approach if we are interested in estimating the association of endogenous TVCs with each cause (Huang et al., 2011; Rizopoulos, 2012).

In the same line of possible applications related to credit, we can design a joint model for stress testing. Stress testing reckons the model's estimations in scenarios that could lead to significant losses and are part of the Basel Accords requirements if the Bank wants to use internal models (BCBS, 2004). In this context, we can think of a joint model that estimates the probabilities of default where the

[2]https://inlabru-org.github.io/inlabru/.

evolution of the longitudinal processes is pushed to extreme scenarios. Moreover, the joint modelling approach has no problem including exogenous TVCs, such as macroeconomic variables, which are relevant when designing adverse economic conditions (Wang et al., 2020). The estimation of the joint model with exogenous TVCs follows similarly as already described in Section 2.2 with the exception that $\boldsymbol{z}_i$ now depends on time. Finally, as we mentioned in Section 4.5, lenders offer various products whose performances are known over time. The use of these multiple sources of data is beneficial for prediction. However, when the frequency of these records is different between products, the data are commonly aggregated, losing valuable information. That is the case with transactional data used for monthly timescale models. Nevertheless, the joint modelling approach does not restrict us from using different time scales between both processes. Therefore, we could eventually link information with a higher granularity in the monthly prediction.

Appendix A

Discrete-Time Joint Model with Autoregressive Terms

A.1 Estimation of Cox model for joint model simulated data

We use the largest simulated data detailed in Section 3.3 (10,000 subjects) and estimate a Cox model where the longitudinal outcome is included as observed (see Table 3.4). We sampled from 3 independent chains with overdispersed starting points, each with 4,000 and 2,000 iterations for the warm-up and sampling periods. Following the same general diagnosis procedure described in Section 3.3 concerning the NUTS sampler, no problems were detected. Table A.1 summarises the parameter estimations. For this model specification under these simulated data, we observed that the 5%-95% credible intervals do not include the true parameter value.

		M_0			
	True	Mean	SD	5%	95%
β_{12}	2.00	1.780	0.061	1.681	1.878
β_{22}	1.00	0.836	0.049	0.756	0.918
λ_f	1.00	0.752	0.024	0.713	0.792

Table A.1: Estimations of M_0 (Cox) for the largest simulated sample.

A.2 Comparing simulations with and without autoregressive term

We are interested in quantifying the relevance of adding at least one autoregressive term in the longitudinal outcome compared to the case with no autoregressive terms. To do so, we perform two simulations analysis. The first one uses the same simulated data from Section 3.3 (10,000 borrowers) and estimates a joint model without the autoregressive term ($\phi = 0$), which is analogous to the specification M_3 in the empirical analysis. We call this model $\widetilde{M}_3$. The second one simulates data as if generated by a joint model without an autoregressive term and estimates a joint model with an autoregressive term. We call this model $\widetilde{M}_5$. The results are shown in Table A.2. We observed that for both models, despite being misspecified for the data, the 5%-95% credible intervals of the parameters related to the covariates of the event process include the true parameters. The differences come from the longitudinal part. We observed that $\widetilde{M}_3$ tries to compensate for misspecification overestimating the fixed effect β_{01}, the variability of the random effects (σ_{U_0} and σ_{U_1}) and the variability of the error terms (σ). However, when the joint model generates the data without an autoregressive term, and we estimate a joint model with an autoregressive term ($\widetilde{M}_5$), we observe that the parameters related to the longitudinal outcome are closer to the true values.

	$\widetilde{M}_3$ with data from $\widetilde{M}_5$					$\widetilde{M}_5$ with data from $\widetilde{M}_3$				
	True	Mean	SD	5%	95%	True	Mean	SD	5%	95%
β_{12}	2.00	1.989	0.070	1.873	2.107	2.00	2.086	0.084	1.949	2.225
β_{22}	1.00	0.937	0.054	0.849	1.026	1.00	1.114	0.070	1.000	1.230
λ_f	1.00	0.990	0.032	0.937	1.043	1.00	1.021	0.051	0.939	1.105
β_{01}	-0.30	-0.473	0.019	-0.505	-0.442	-0.30	-0.291	0.012	-0.311	-0.271
ϕ	0.40					0.00	0.008	0.002	0.004	0.011
σ	1.00	1.044	0.002	1.042	1.047	1.00	1.002	0.001	1.000	1.005
σ_{U_0}	1.20	1.945	0.014	1.921	1.969	1.20	1.168	0.010	1.152	1.184
σ_{U_1}	0.05	0.090	0.001	0.089	0.092	0.05	0.050	0.001	0.049	0.051
ρ_U	-0.20	-0.196	0.011	-0.215	-0.177	-0.20	-0.175	0.013	-0.196	-0.154

Table A.2: Estimations of $\widetilde{M}_3$ (joint model without AR1) for data coming from $\widetilde{M}_5$ (left) and estimations of $\widetilde{M}_5$ (joint model with AR1) for data coming from $\widetilde{M}_3$ (right).

A.3 Survival probability ranges

Table A.3 shows the probability ranges (5-95%) for non-defaulters (value 0) and defaulters (value 1) for the 6 estimated models.

	M_0		M_1		M_2		M_3		M_4		M_5	
Time(c)	0	1	0	1	0	1	0	1	0	1	0	1
6	0.95-1.00	0.90-1.00	0.98-1.00	0.96-1.00	0.98-1.00	0.96-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.94-1.00
7	0.95-1.00	0.89-1.00	0.97-1.00	0.95-1.00	0.98-1.00	0.95-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.94-1.00
8	0.97-1.00	0.92-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.93-1.00
9	0.97-1.00	0.93-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.93-1.00	0.98-1.00	0.94-1.00	0.97-1.00	0.93-1.00
10	0.97-1.00	0.92-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.94-1.00	0.95-1.00	0.92-1.00	0.98-1.00	0.95-1.00	0.97-1.00	0.94-1.00
11	0.96-1.00	0.91-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.94-1.00	0.92-1.00	0.80-1.00	0.97-1.00	0.92-1.00	0.97-1.00	0.93-1.00
12	0.96-1.00	0.91-1.00	0.96-1.00	0.93-1.00	0.96-1.00	0.93-1.00	0.87-1.00	0.66-1.00	0.94-1.00	0.86-1.00	0.96-1.00	0.93-0.99
13	0.96-1.00	0.91-1.00	0.96-1.00	0.93-1.00	0.96-1.00	0.93-1.00	0.80-1.00	0.51-1.00	0.87-1.00	0.67-1.00	0.95-1.00	0.89-0.99
14	0.96-1.00	0.90-1.00	0.96-1.00	0.92-0.99	0.96-1.00	0.92-0.99	0.74-1.00	0.39-1.00	0.73-1.00	0.43-1.00	0.94-1.00	0.85-0.99
15	0.95-1.00	0.90-1.00	0.96-1.00	0.92-0.99	0.96-1.00	0.92-0.99	0.70-1.00	0.21-1.00	0.55-1.00	0.12-1.00	0.92-1.00	0.77-0.99
16	0.95-1.00	0.89-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.67-1.00	0.18-1.00	0.37-1.00	0.04-1.00	0.90-1.00	0.71-0.99
17	0.95-1.00	0.89-1.00	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.66-1.00	0.16-1.00	0.24-1.00	0.01-1.00	0.87-1.00	0.62-1.00
18	0.95-1.00	0.90-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.67-1.00	0.20-1.00	0.17-1.00	0.01-1.00	0.85-1.00	0.62-1.00
19	0.95-1.00	0.91-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.69-1.00	0.21-1.00	0.14-1.00	0.00-1.00	0.84-1.00	0.56-0.99
20	0.95-1.00	0.91-1.00	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.71-1.00	0.37-0.99	0.13-1.00	0.01-1.00	0.83-1.00	0.59-0.99
21	0.95-1.00	0.91-0.99	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.74-1.00	0.37-0.99	0.14-1.00	0.01-1.00	0.83-1.00	0.58-0.99
22	0.95-1.00	0.90-1.00	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.77-1.00	0.41-1.00	0.16-1.00	0.01-1.00	0.83-1.00	0.58-1.00
23	0.95-1.00	0.91-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.78-1.00	0.45-1.00	0.19-1.00	0.02-1.00	0.83-1.00	0.59-1.00
24	0.95-1.00	0.90-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.79-1.00	0.52-1.00	0.23-1.00	0.03-1.00	0.83-1.00	0.62-1.00

Table A.3: Survival probability ranges (5-95%) for non-defaulters (value 0) and defaulters (value 1) (see $\hat{\pi}_k(c+12|c)$ in Equation 3.11). The Time(c) column represents c, the known history of the subjects.

A.4 Calibration sensitivity analysis

Our interest is to investigate how sensitive the calibration of the joint model M_5 is to the class imbalance in comparison to the benchmark. To this end, we perform a five-fold cross-validation analysis similar to the one described in Section 3.4. Still, we now randomly reduce the non-defaulters proportion in the training folds (down-sampling). We perform the analysis for two different non-defaulters proportions, one corresponding to 75% of the loans and the other to 50%. Table A.4 shows the mean differences and standard deviations of PE with respect to M_0 for the range of c and the forecast window of 12 months for both class proportions. Although we observe that both models, M_0 and M_5, are sensible to class imbalance showing improvements in their calibration when compared to the results shown in Table 3.7, the joint model has reasonably decreased the difference in the PE ($\Delta\widehat{PE}_c^{12}\ M_5$), especially for $c \geq 15$ where the most significant differences were observed before.

A.5 Robustness checks

To study the robustness of the results shown in Table 3.5, we re-estimate the model that has the most complex structure, M_5, using different priors. We keep the noninformative uniform priors for λ_f, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ϕ. Moreover, for the covariance matrix Σ, we set the scale parameter of the LKJ distribution to 1, which corresponds to the uniform density over correlation matrices. In addition, for both variability terms, σ and $\theta_{\boldsymbol{\alpha}}$, instead of using a uniform and a half-Cauchy priors, respectively, we use for both the inverse Gamma with shape 1 and scale 0.001, as suggested by Ibrahim et al. (2001, Ch.7). Recall that σ is the standard deviation of the error terms (Equation 3.1) and $\theta_{\boldsymbol{\alpha}}$ is the hyperparameter of the vector of B-spline coefficients $\boldsymbol{\alpha}$ (Equation 3.8). That is to say, instead of assuming $\theta_{\boldsymbol{\alpha}} \sim$ half-Cauchy(25) for $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \theta_{\boldsymbol{\alpha}}^2 I)$, we assume $\theta_{\boldsymbol{\alpha}} \sim$ inverse-Gamma(1, 0.001). To illustrate how different these two distributions are, Table A.5 shows various percentiles for each of them.

In Table A.6, under the name M_5 and to facilitate comparison, we show again the results of the parameters associated with model M_5 from Table 3.5. Alos, the results obtained using these new priors are found under the name of $\tilde{M}_5$. We observe that the parameter estimates remain consistent when using these different

	25%-75%		50%-50%	
Time(c)	$\widehat{PE}_c^{12}$ M_0	$\Delta\widehat{PE}_c^{12}$ M_5	$\widehat{PE}_c^{12}$ M_0	$\Delta\widehat{PE}_c^{12}$ M_5
6	0.354	**-0.010 (0.006)**	0.348	**-0.005 (0.003)**
7	0.384	**-0.009 (0.006)**	0.379	**-0.004 (0.003)**
8	0.418	**-0.007 (0.006)**	0.414	**-0.004 (0.004)**
9	0.458	**-0.005 (0.004)**	0.455	**-0.003 (0.002)**
10	0.481	**-0.004 (0.003)**	0.478	**-0.002 (0.002)**
11	0.523	**-0.003 (0.002)**	0.521	-0.001 (0.001)
12	0.584	0.000 (0.002)	0.582	0.002 (0.000)
13	0.611	0.005 (0.004)	0.609	0.007 (0.003)
14	0.673	0.029 (0.030)	0.670	0.021 (0.010)
15	0.737	0.067 (0.083)	0.734	0.040 (0.030)
16	0.797	0.110 (0.152)	0.796	0.061 (0.053)
17	0.798	0.165 (0.212)	0.797	0.089 (0.074)
18	0.845	0.217 (0.263)	0.842	0.117 (0.092)
19	0.905	0.237 (0.284)	0.903	0.124 (0.099)
20	0.912	0.255 (0.289)	0.910	0.132 (0.103)
21	0.946	0.250 (0.243)	0.946	0.131 (0.081)
22	0.913	0.220 (0.239)	0.914	0.110 (0.085)
23	0.914	0.184 (0.143)	0.914	0.099 (0.043)
24	0.904	0.200 (0.153)	0.905	0.110 (0.055)

***For ease of visualisation, all values are multiplied by 100.**

Table A.4: Mean difference of $\widehat{PE}_c^{\Delta c}$ (Equation 3.14) between models M_5 and M_0 for prediction window of 12 months ($\Delta c = 12$) considering two down-sampling settings; 75% and 50% of non-defaulters. The Time(c) column represents c, the known history when making the prediction. The number in parentheses is the standard deviation of the cross-validation analysis.

prior distributions.

Likewise, in Table A.7 we show the results for the hyperparameter, $\theta_{\boldsymbol{\alpha}}$, and the B-spline coefficients $\boldsymbol{\alpha}$. We can see that although there are differences between the estimates of the hyperparameter $\theta_{\boldsymbol{\alpha}}$ when using both priors, the results corresponding to the B-splines coefficients remain practically the same, which agrees with the results in Table A.6.

	10%	25%	50%	75%	90%
Half-Cauchy(25)	3.95961	10.35534	25.00000	60.35534	157.84379
Inverse-Gamma(1, 0.001)	0.00043	0.00071	0.00143	0.00344	0.00941

Table A.5: Comparison of percentiles between half-Cauchy with a scale of 25 and an inverse-Gamma with shape 1 and scale 0.001.

Parameter	M_5			$\tilde{M}_5$		
	Mean	5%	95%	Mean	5%	95%
fico	-0.701	-0.821	-0.581	-0.699	-0.820	-0.579
cltv	0.516	0.333	0.703	0.514	0.339	0.703
orig_upb	-0.155	-0.300	-0.014	-0.154	-0.304	-0.012
dti	0.152	0.021	0.283	0.150	0.022	0.283
n_borr	-0.270	-0.527	-0.018	-0.277	-0.528	-0.029
loan_purpose	-0.971	-1.246	-0.696	-0.970	-1.243	-0.706
λ_f	1.317	0.895	1.771	1.316	0.901	1.750
β_{01}	-0.280	-0.294	-0.266	-0.280	-0.293	-0.268
σ_{U_0}	1.237	1.219	1.255	1.237	1.218	1.256
σ	0.706	0.704	0.708	0.706	0.704	0.708
ϕ	0.357	0.353	0.360	0.357	0.354	0.360
σ_{U_1}	0.053	0.052	0.054	0.053	0.052	0.054
ρ_U	-0.811	-0.818	-0.804	-0.811	-0.819	-0.803

Table A.6: Summary of parameter estimates of the model M_5 using different prior distributions and with fold one kept out. To ease comparison, the three columns below M_5 are copied from Table 3.5 and the three below $\tilde{M}_5$ are the new results.

Parameter	M_5			$\tilde{M}_5$		
	Mean	5%	95%	Mean	5%	95%
$\theta_{\boldsymbol{\alpha}}$	8.344	5.206	13.371	7.122	4.609	11.034
α_1	-8.577	-9.605	-7.655	-8.534	-9.584	-7.638
α_2	-8.047	-9.227	-6.880	-8.064	-9.193	-6.948
α_3	-6.583	-7.475	-5.695	-6.550	-7.401	-5.713
α_4	-6.245	-6.877	-5.617	-6.252	-6.870	-5.624
α_5	-6.052	-6.893	-5.239	-6.051	-6.884	-5.244
α_6	-6.209	-7.098	-5.350	-6.191	-7.053	-5.372
α_7	-6.315	-7.051	-5.632	-6.318	-7.064	-5.657

Table A.7: Parameter estimates associated with the vector of B-spline functions of the model M_5 using different prior distributions and with fold one kept out.

Appendix B

Joint Model of Multivariate Longitudinal Outcome

B.1 Comparison between MCMC and INLA estimations

This section aims to illustrate how fast and accurate the INLA methodology is compared to an MCMC sampling scheme for the multivariate joint model presented in Chapter 4. To this extent, we implement the multivariate joint model using the platform for statistical modelling *Stan* with the No-U-Turn Sampler (NUTS Hoffman and Gelman, 2014), which is regarded as a faster extension to Hamiltonian Monte Carlo algorithm. To assess convergence of the NUTS sampler, we performed the sampling from 3 independent chains with overdispersed starting points and, following the general diagnosis detailed in Betancourt (2017), no convergence problems were detected.

For the simulation setting described in Section 4.3, we estimate the model via MCMC and INLA using the same computational resources (6 CPU cores, each with 4 GB of memory). We measure the times each procedure takes for different numbers of simulated loans (N), ranging from 250 to 500. The times, in minutes, are shown in Table B.1. We can observe that considering, for example, a sample with 300 simulated loans, which is relatively small, the time required by the MCMC estimation is more than 2 hours, whereas for the INLA version is less than 2 minutes.

N	$T_{\text{MCMC}}(\text{min})$	$T_{\text{INLA}}(\text{min})$
250	106.03	1.07
300	127.79	1.31
350	159.52	1.50
400	187.70	1.99
450	207.00	2.00
500	256.12	2.42

Table B.1: Time required, in minutes, for model estimation through MCMC and INLA schemes as a function of the number of loans (N).

Moreover, for the biggest sample size simulated in this analysis ($N = 500$), we save the marginal posterior distributions for each parameter in the simulation setting (see Section 4.3) and estimate their credible intervals obtained by both implementations. The comparison of the $2.5\% - 97.5\%$ credible intervals are shown in Figure B.1. First, we notice that both implementations estimate intervals that include the true parameter value for all the parameters in the simulation setting. Second, both are fairly similar, evidencing and supporting the quality of the Bayesian inference approximation performed by INLA for our model specification.

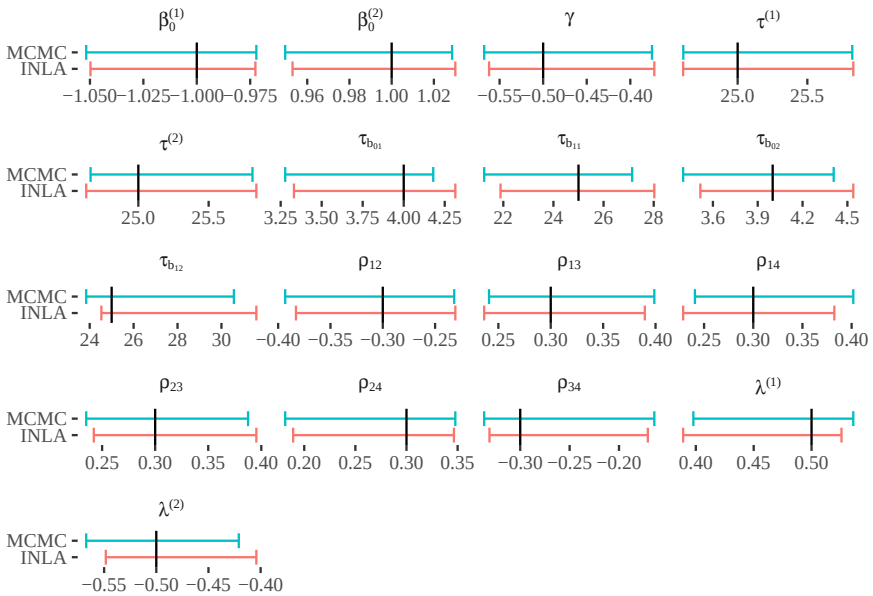

Figure B.1: Credible intervals ($2.5\% - 97.5\%$) obtained by the MCMC and INLA implementations for each parameter in the simulation analysis. The solid vertical line corresponds to the true parameter value.

B.2 Time-fixed covariates distributions

Figure B.2 shows the distribution for the six fixed covariates provided: four continuous and two categorical. Due to data confidentiality agreements, not all covariates can be named, and the x-axes of the plots are omitted. We also show the parameter estimate sign next to the covariate name in parentheses. The signs are consistent among all the estimated models. Note that the effect of age and debt-to-income is negative for the probability of prepayment. However, the effect is positive for the interest rate given at origination and the loan amount.

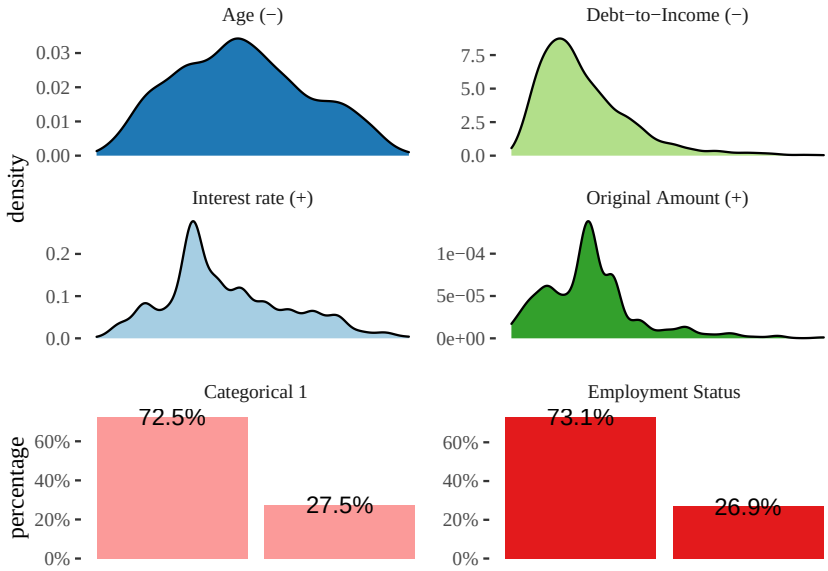

Figure B.2: Distribution of the time-fixed covariates included in the survival model. For the bank privacy concerns, some information is omitted. The sign in parentheses is the sign of the parameters estimates.

Appendix C

Spatio-Temporal Joint Models

C.1 Estimation of cvDCL under MCMC scheme

In Section 5.2.3 we show how the *cross-validated Dynamic Conditional Likelihood* (cvDCL) is estimated using the INLA methodology. Here, we describe how the cvDCL is estimated with an MCMC sampling scheme. This is done for the sake of completeness since in Appendix C.2 we compare numerically how different these two approaches are using simulation analysis.

Recall from Equation 5.8 that the cvDCL is defined as

$$\text{cvDCL}(t) = \frac{1}{N_t}\sum_{i=1}^{N} -I(T_i > t)\log\{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})\},$$

where $N_t = \sum_{i=1}^{N} I(T_i > t)$ (the number of loans at risk at time t). It can be shown that (see Rizopoulos et al., 2016, for further details)[1]

$$p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})^{-1} \approx \int \frac{p(\boldsymbol{U}_i, \boldsymbol{\Theta}|\mathcal{D})}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{U}_i, \boldsymbol{\Theta})} \mathrm{d}\boldsymbol{\Theta}\mathrm{d}\boldsymbol{U}_i, \qquad \text{(C.1)}$$

where $\boldsymbol{\Theta}$ is the set of all parameters as described in Section 5.2.2 and $\boldsymbol{U}_i$ the random effects for loan i. Let $\boldsymbol{\Theta}^{(g)}$ and $\boldsymbol{U}_i^{(g)}$ denote the gth realisation of the posterior sample with $g = 1, ..., G$, then $p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})^{-1}$ can be estimated by

$$\hat{p}(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \mathcal{D}_{-i})^{-1} = \frac{1}{G}\sum_{g=1}^{G} \frac{1}{p(T_i, \delta_i | T_i > t, \boldsymbol{y}_i(t), \boldsymbol{U}_i^{(g)}, \boldsymbol{\Theta}^{(g)})}.$$

[1] In that work, Equation C.1 is presented as an equality. We confirmed with the author that there is an error and that it should be an approximation symbol instead.

Hence, cvDCL(t) can be estimated as

$$\widehat{\text{cvDCL}}(t)^{MCMC} = \frac{1}{N_t}\sum_{i=1}^{N} I(T_i > t)\log\left\{\frac{1}{G}\sum_{g=1}^{G}\frac{1}{p(T_i,\delta_i|T_i>t,\boldsymbol{y}_i(t),\boldsymbol{U}_i^{(g)},\boldsymbol{\Theta}^{(g)})}\right\}. \tag{C.2}$$

We estimate the variance of $\widehat{\text{cvDCL}}(t)^{MCMC}$ through batching (Carlin and Louis, 2000). This requires that a long run of G samples is divided into M successive batches of size H (i.e. $G = M \cdot H$). For each batch $m = 1, ..., M$, we calculate $\widehat{\text{cvDCL}}(t)_m^{MCMC}$ using its H posterior samples and the variance is then the sample variance of these M estimations.

C.2 Comparison cvDCL: MCMC and INLA

In this appendix, we analyse how different is the estimation of the cvDCL between the MCMC and INLA procedures (Equations C.2 from Appendix C.1 and 5.10 from Section 5.2.3, respectively). To this end we, first, generate data from a joint model that follows Equations C.3 and C.4 for the longitudinal and event processes, respectively.

$$\begin{aligned}(Y_{i,s}|\eta_{Yi,s},\tau_Y) &\sim N(\eta_{Yi,s},\tau_Y^{-1})\\ \eta_{Yi,s} &= \beta_{01} + U_{0i} + (\beta_{11} + U_{1i})s,\\ (U_{0i},U_{1i})^\intercal &\sim N_2(\mathbf{0},Q_{\mathbf{U}}^{-1}),\end{aligned} \tag{C.3}$$

$$\begin{aligned}(X_{i,s}|X_{i,s-1}=0,\eta_{Xi,s}) &\sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{Xi,s}))\\ \eta_{Xi,s} &= \nu_0 + v_s + \beta_{12}z_{1i} + \beta_{22}z_{2i} + \lambda(U_{0i} + U_{1i}s),\\ v_s &\sim RW2(\tau_v).\end{aligned} \tag{C.4}$$

Next, we estimate the $\widehat{\text{cvDCL}}(t)^{MCMC}$ and $\widehat{\text{cvDCL}}(t)^{INLA}$, for different values of t, assuming two different specifications of the joint model. The first specification is the correct one, i.e. follows Equations C.3 and C.4. The second one omits the second covariate in the event predictor, specifically, it is assumed that the linear predictor of the event process is $\nu_0 + v_s + \beta_{11}z_{1i} + \lambda(U_{0i} + U_{1i}s)$ (see Equation C.4). This analysis not only allows us to compare the values of the cvDCL for two distinct settings but also help us to measure how different the cvDCL is when one specification explains the data better than the other one.

Table C.1 shows the results of the comparative analysis.

		Correct Specification		
	N_t	MCMC	INLA Lap	INLA EB
$\widehat{\text{cvDCL}}(t=12)$	424	2.5915 (6.80e-04)	2.5922	2.5778
$\widehat{\text{cvDCL}}(t=18)$	347	2.4715 (6.75e-04)	2.4716	2.4679
$\widehat{\text{cvDCL}}(t=24)$	183	2.3951 (8.66e-04)	2.3942	2.3930
$\widehat{\text{cvDCL}}(t=30)$	85	2.1884 (1.63e-03)	2.1857	2.1864
$\widehat{\text{cvDCL}}(t=36)$	36	1.7153 (4.22e-03)	1.7085	1.7115
		Other Specification		
	N_t	MCMC	INLA Lap	INLA EB
$\widehat{\text{cvDCL}}(t=12)$	424	2.8023 (7.34e-04)	2.8027	2.7949
$\widehat{\text{cvDCL}}(t=18)$	347	2.6893 (7.95e-04)	2.6890	2.6874
$\widehat{\text{cvDCL}}(t=24)$	183	2.6037 (1.27e-03)	2.6027	2.6031
$\widehat{\text{cvDCL}}(t=30)$	85	2.3634 (1.70e-03)	2.3598	2.3632
$\widehat{\text{cvDCL}}(t=36)$	36	1.8390 (4.27e-03)	1.8305	1.8382

Table C.1: Comparison of model performance for simulated data and two different specifications.

Bibliography

Albert, P. S. and Shih, J. H. (2010a). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, 4(3):1517.

Albert, P. S. and Shih, J. H. (2010b). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*, 66(3):983–987.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13(1982):61–98.

Alsefri, M., Sudell, M., García-Fiñana, M., and Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: a methodological review. *BMC Medical Research Methodology*, 20:1–17.

Andrinopoulou, E.-R., Eilers, P. H., Takkenberg, J. J., and Rizopoulos, D. (2018). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using p-splines. *Biometrics*, 74(2):685–693.

Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J., and Lesaffre, E. (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in medicine*, 33(18):3167–3178.

Arminger, G., Enache, D., and Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. *Computational Statistics*, 12(2).

Bacci, S., Bartolucci, F., and Pandolfi, S. (2018). A joint model for longitudinal and survival data based on an ar (1) latent process. *Statistical Methods in Medical Research*, 27(5):1285–1311.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.

Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science*, 49(3):312–329.

Banasik, J., Crook, J. N., and Thomas, L. C. (1999). Not if but when will

borrowers default. *Journal of the Operational Research Society*, 50(12):1185–1190.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Barrett, J., Diggle, P., Henderson, R., and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 77(1):131.

BCBS (2000). Principles for the management of credit risk. Technical report, Bank for International Settlements.

BCBS (2004). International convergence of capital measurement and capital standards: a revised framework. Technical report, Bank for International Settlements.

BCBS (2019). The Basel Framework. Technical report, Bank for International Settlements.

Beck, T., Levine, R., and Loayza, N. (2000). Finance and the sources of growth. *Journal of financial economics*, 58(1-2):261–300.

Bellotti, T. and Crook, J. (2009a). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60(12):1699–1707.

Bellotti, T. and Crook, J. (2009b). Support vector machines for credit scoring and discovery of significant features. *Expert systems with applications*, 36(2):3302–3308.

Bellotti, T. and Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, 29(4):563–574.

Bellotti, T. and Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society*, 65(3):340–350.

Benavoli, A., Corani, G., Demšar, J., and Zaffalon, M. (2017). Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Bhattacharya, A., Wilson, S. P., and Soyer, R. (2019). A Bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research*, 274(3):1112–1124.

Blumenstock, G., Lessmann, S., and Seow, H.-V. (2022). Deep learning for survival and competing risk modelling. *Journal of the Operational Research Society*, 73(1):26–38.

Bremhorst, V. and Lambert, P. (2016). Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics & Data Analysis*, 93:270–284.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Brown, E. R., Ibrahim, J. G., and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73.

Calabrese, R., Andreeva, G., and Ansell, J. (2019). "birds of a feather" fail together: Exploring the nature of dependency in sme defaults. *Risk Analysis*, 39(1):71–84.

Calabrese, R. and Crook, J. (2020). Spatial contagion in mortgage defaults: A spatial dynamic survival model with time and space varying coefficients. *European Journal of Operational Research*, 287(2):749–761.

Calabrese, R., Marra, G., and Angela Osmetti, S. (2016). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *Journal of the Operational Research Society*, 67(4):604–615.

Calabrese, R. and Osmetti, S. A. (2013). Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics*, 40(6):1172–1188.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC, Boca Raton.

Chang, H. H., Reich, B. J., and Miranda, M. L. (2013). A spatial time-to-event approach for estimating associations between air pollution and preterm birth. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(2):167–179.

Chi, Y.-Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2):432–445.

Clayton, D. G. (1996). Generalized linear mixed models. *Markov chain Monte Carlo in practice*, 1:275–302.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Consalvi, M. and Scotto di Freca, G. (2010). Measuring prepayment risk: an application to UniCredit Family Financing. Technical report, UniCredit and Universities.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Creal, D. D., Gramacy, R. B., and Tsay, R. S. (2014). Market-based credit ratings. *Journal of Business & Economic Statistics*, 32(3):430–444.

Crook, J. and Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):283–305.

Dastile, X., Celik, T., and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263.

Dirick, L., Bellotti, T., Claeskens, G., and Baesens, B. (2019). Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. *Journal of Business & Economic Statistics*, 37(1):40–53.

Djeundje, V. B. and Crook, J. (2018). Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. *European Journal of Operational Research*, 271(2):697–709.

Djeundje, V. B. and Crook, J. (2019a). Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research*, 275(1):319–333.

Djeundje, V. B. and Crook, J. (2019b). Identifying hidden patterns in credit risk survival data using generalised additive models. *European Journal of Operational Research*, 277(1):366–376.

Doumpos, M., Lemonakis, C., Niklis, D., and Zopounidis, C. (2019). *Analytical techniques in the assessment of credit risk*. Springer.

Duffie, D. (2005). Credit risk modeling with affine processes. *Journal of Banking & Finance*, 29(11):2751–2802.

Elashoff, R. M., Li, G., and Li, N. (2016). *Joint Modeling of Longitudinal and Time-to-Event Data*. CRC Press.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, volume 425. Springer, New York, NY.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC press.

Freni-Sterrantino, A., Ventrucci, M., and Rue, H. (2018). A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and spatio-temporal epidemiology*, 26:25–34.

Furgal, A. K., Sen, A., and Taylor, J. M. (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *International Statistical Review*, 87(2):393–418.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Goicoa, T., Adin, A., Ugarte, M., and Hodges, J. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment*, 32(3):749–770.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

Goodstein, R., Hanouna, P., Ramirez, C. D., and Stahel, C. W. (2017). Contagion effects in strategic mortgage defaults. *Journal of Financial Intermediation*, 30:50–60.

Gross, D. B. and Souleles, N. S. (2002). An empirical analysis of personal bankruptcy and delinquency. *The Review of Financial Studies*, 15(1):319–347.

Guiso, L., Sapienza, P., and Zingales, L. (2013). The determinants of attitudes toward strategic default on mortgages. *The Journal of Finance*, 68(4):1473–1515.

Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., and Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1):292–305.

Gupta, A. (2019). Foreclosure contagion and the neighborhood spillover effects of mortgage defaults. *The Journal of Finance*, 74(5):2249–2301.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press, Cambridge, MA.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*, volume 451. John Wiley & Sons.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Henderson, R., Diggle, P., and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics*, 3(1):33–50.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC medical research methodology*, 16(1):117.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hu, W. and Zhou, J. (2019). Joint modeling: an application in behavioural scoring. *Journal of the Operational Research Society*, 70(7):1129–1139.

Huang, X., Li, G., Elashoff, R. M., and Pan, J. (2011). A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime data analysis*, 17(1):80–100.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*, volume 2. Springer Science & Business Media, New York.

Jaffa, M. A., Gebregziabher, M., and Jaffa, A. A. (2014). A joint modeling approach for right censored high dimensional multivariate longitudinal data. *Journal of Biometrics & Biostatistics*, 5(4).

Jaffa, M. A., Woolson, R. F., and Lipsitz, S. R. (2011). Slope estimation for bivariate longitudinal outcomes adjusting for informative right censoring by using a discrete survival model: application to the renal transplant cohort. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):387–402.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9):6233–6239.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis: A Self-Learning Text*. Springer, New York, NY.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, 19(17-18):2555–2567.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lawrence Gould, A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., and Bois, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine*, 34(14):2181–2195.

Leow, M. and Crook, J. (2014). Intensity models and transition probabilities for credit card loan delinquencies. *European Journal of Operational Research*, 236(2):685–694.

Leow, M. and Crook, J. (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research*, 249(2):457–464.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian journal of statistics*, 35(4):691–700.

Liu, W., Chase, J. M., Vu, C., Cela, J., et al. (2009). Generalizations of generalized additive model (gam): A case of credit risk modeling. *SAS Global*.

Ma, P., Crook, J., and Ansell, J. (2010). Modelling take-up and profitability. *Journal of the Operational Research Society*, 61(3):430–442.

Martins, R., Silva, G. L., and Andreozzi, V. (2016). Bayesian joint modeling of longitudinal and spatial survival AIDS data. *Statistics in medicine*, 35(19):3368–3384.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67:68–83.

McDonald, R. A., Matuszyk, A., and Thomas, L. C. (2010). Application of survival analysis to cash flow modelling for mortgage products. *OR insight*, 23(1):1–14.

Medina-Olivares, V., Calabrese, R., Crook, J., and Lindgren, F. (Accepted/in press 2022a). Joint models for longitudinal and discrete survival data in credit scoring. *European Journal of Operational Research*.

Medina-Olivares, V., Calabrese, R., Dong, Y., and Shi, B. (2022b). Spatial dependence in microfinance credit default. *International Journal of Forecasting*, 38(3):1071–1085.

Nadeau, C. and Bengio, Y. (2000). Inference for the generalization error. In *Advances in Neural Information Processing Systems*, pages 307–313.

Narain, B. (1992). Survival analysis and the credit granting decision. *Credit scoring and credit control*, 109:121.

Papageorgiou, G. and Rizopoulos, D. (2021). An alternative characterization of MAR in shared parameter models for incomplete longitudinal data and its utilization for sensitivity analysis. *Statistical Modelling*, 21(1-2):95–114.

Pence, K. M. (2006). Foreclosing on opportunity: State laws and mortgage credit. *Review of Economics and Statistics*, 88(1):177–182.

Pennington-Cross, A. (2010). The duration of foreclosures in the subprime mortgage market: a competing risks model with mixing. *The Journal of Real Estate Finance and Economics*, 40(2):109–129.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, 10(3):535–549.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ratcliffe, S. J., Guo, W., and Ten Have, T. R. (2004). Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*, 60(4):892–899.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12):1366–1380.

Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association*, 109(508):1385–1397.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276.

Rizopoulos, D., Taylor, J. M., Van Rosmalen, J., Steyerberg, E. W., and Takkenberg, J. J. (2016). Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1):149–164.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics*, 22(6):725–734.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, 74(1):101–124.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, pages 1–28.

Stan Development Team (2018). Cmdstan: The command-line interface to stan. `http://mc-stan.org`.

Stan Development Team and others (2022). Stan modeling language user's guide and reference manual, version 2.29. 0.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, 50(2):277–289.

Stepanova, M. and Thomas, L. C. (2001). PHAB scores: proportional hazards analysis behavioural scores. *Journal of the Operational Research Society*, 52(9):1007–1016.

Thomas, L., Crook, J., and Edelman, D. (2017). *Credit Scoring and its Applications*, volume 2. Society for Industrial and Applied Mathematics.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Towe, C. and Lawley, C. (2013). The contagion effect of neighboring foreclosures. *American Economic Journal: Economic Policy*, 5(2):313–35.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

Tutz, G. and Schmid, M. (2016). *Modeling discrete time-to-event data*. Springer, New York.

Van Niekerk, J., Bakka, H., and Rue, H. (2019). Joint models as latent gaussian models-not reinventing the wheel. *arXiv preprint arXiv:1901.09365*.

Ver Hoef, J. M. (2012). Who invented the delta method? *The American Statistician*, 66(2):124–127.

Wang, Y., Wang, S., and Lai, K. K. (2005). A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, 13(6):820–831.

Wang, Z., Crook, J., and Andreeva, G. (2020). Reducing estimation risk using a Bayesian posterior distribution approach: Application to stress testing mortgage loan default. *European Journal of Operational Research*, 287(2):725–738.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 64(327):1073–1078.

Wu, L. (2009). *Mixed effects models for complex data*. Chapman and Hall/CRC press.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339.

Zhou, H., Lawson, A. B., Hebert, J. R., Slate, E. H., and Hill, E. G. (2008). Joint spatial survival modeling for the age at diagnosis and the vital outcome of prostate cancer. *Statistics in medicine*, 27(18):3612–3628.